Gene Cloning and Manipulation

Second Edition

Christopher Howe
University of Cambridge

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK
Published in the United States of America by Cambridge University Press, New York
www.cambridge.org
Information on this title: www.cambridge.org/9780521817936

First published in print format 2007

ISBN-13 978-0-511-29499-0 eBook (EBL)
ISBN-10 0-511-29499-9 eBook (EBL)

ISBN-13 978-0-521-81793-6 hardback
ISBN-10 0-521-81793-5 hardback

ISBN-13 978-0-521-52105-5 paperback
ISBN-10 0-521-52105-X paperback

Contents

Preface to first edition

This book grew out of sets of lectures given to undergraduates taking courses in Biochemistry and Molecular Biology, and Medical Sciences. I hope it will be useful to people studying a range of biological subjects. I have tried to concentrate throughout on the general principles underlying the subject rather than to give overwhelmingly detailed accounts of vector systems and practical instructions. For those, there are more detailed books, reviews, catalogues, and lab manuals. I am grateful to the many friends and colleagues who have helped in the production of this book by reading sections (and in some cases the whole thing!) or in other ways. In particular, I should like to thank Janet Allen, Alison Baker, Adrian Barbrook, Alison Franklin, Hilary and Tony Larkum, and Saul Purton. I am also grateful to Robin Smith of Cambridge University Press for his advice and encouragement throughout the exercise, and to Robert Sugar and Dorothy Duncan of Bookworks for their help in the book's production.

Christopher Howe

Preface to second edition

The field of gene cloning and manipulation has changed dramatically since the first edition of this book appeared, and this development is reflected in the changes I have introduced in the second edition. The applications of PCR methods have expanded enormously, and ''omics'' and reverse genetic technologies are available across a wide range of organisms. Significant improvements have also been made in established areas, such as in the hosts and vectors for protein expression, and in the use of fluorescent proteins as reporter genes. As with the first edition, I have tried to stress the principles underlying the vectors we use, and avoid long and detailed lists (which would soon become out of date, anyway). Recognizing the necessity of being able to devise appropriate strategies for individual experimental situations, I have added a final chapter that gives examples and suggestions.

I am grateful to the members of my lab who waited patiently while the pressure of finishing this edition (which became known as my 'big book of fun'!) delayed other things. I am particularly grateful to the people who helped directly in various ways, especially Mim Bower, Jon Burton, Ellen Nisbet, Saul Purton, Beatrix Schlarb-Ridley, and Petrus de Vries. I would also like to thank Katrina Halliday and Clare Georgy of Cambridge University Press, together with Peter Lewis and Rasika Mathur of Keyword Group for their technical expertise, patience and encouragement.

Christopher Howe

Chapter 1

The tools for the job

1.1 Introduction

Cloning and manipulating genes requires the ability to cut, modify and join genetic material (usually DNA, but sometimes RNA) and check the parameters of the molecules, such as size, that are being manipulated. We will assume knowledge of the structure of the materials involved (DNA, RNA and so on) and start by describing the tools available for manipulating them. Many of the tools involved are enzymes that have important physiological roles in cells. To understand why they are useful for our purposes, we should be aware of their normal roles, too.

The choice of which enzyme is used for a particular purpose depends mainly on two considerations:

1. How easy (i.e. inexpensive) is it to purify? This will be determined by its abundance in the cell and by how easy it is to separate it from other undesirable activities.
2. How well does it do the job? This will depend upon its specificity ('accuracy') and specific activity ('speed') and upon the details of the reaction which it catalyses.

Other factors, such as stability, are also important.

Techniques of genetic manipulation can be applied to the production of the enzymes for genetic manipulation itself. It is possible to use cloned genes to prepare large quantities of these enzymes more easily, as well as to modify the genes to 'improve' their function, perhaps by slightly altering the properties of the enzymes they encode.

1.2 Cutting

Enzymes that break down nucleic acids are called nucleases. Those that break down RNA are called ribonucleases, or RNases, and those that break down DNA are called deoxyribonucleases or DNases.

There are two ways of breaking down a linear nucleic acid molecule: dismantling it bit by bit from the ends, or breaking it into pieces by cutting within the molecule. The former is called exonucleolytic activity (Greek *exo* = outside) and the latter endonucleolytic activity (Greek *endon* = within). Do not fall into the trap of thinking that *endo*nucleases work from the *end*s in! For cutting nucleic acid molecules into pieces, therefore, we will need endonucleases, and the most widely used ones are the restriction endonucleases.

1.2.1 Restriction endonucleases

Restriction endonucleases are part of the natural defence mechanisms of bacteria against incoming DNA, which may be from viruses or plasmids from a foreign population of cells. These enzymes were first recognized by their ability to restrict the growth of certain viruses in particular strains of *Escherichia coli*, and were named accordingly. (The verb restrict is now widely used by molecular biologists to mean 'cut with a restriction endonuclease'.) The restriction enzymes are associated with modifying enzymes, which methylate the DNA. Methylation protects the DNA from cleavage by endonucleases, and this stops the cell from degrading its own DNA. Invading DNA that has not been correctly methylated will be degraded unless it can be modified by the cell's methylating enzymes quickly enough, which happens only rarely.

DNA, once modified, remains protected even after replication. This is because semiconservative replication of a molecule methylated on both strands results in two daughter molecules that are hemimethylated (i.e. methylated on one strand), and hemimethylation is sufficient to confer protection against cleavage by an endonuclease. The non-methylated strand can then be modified before replication takes place again.

Three types of restriction/modification system are recognized. These are called Types (or Classes) I, II and III, and their key properties are summarized in Table 1.1. All the enzymes recognize particular DNA sequences, but only the Type II endonucleases cut within those recognition sequences. The recognition sites for a number of Type II enzymes are given in Table 1.2. These enzymes often make a 'staggered' cut to leave molecules that, although primarily double stranded, have short single-stranded ends. These are called sticky ends. Depending on the enzyme, either the 5′ end or the 3′ end may be left single stranded. The molecules generated have a phosphate group on the 5′ end and a hydroxyl group on the 3′ end. A small number of Type II enzymes cut just outside their recognition sites; for example, *Mbo*II cuts seven nucleotides 3′ to its recognition site of -GAAGA-. Others cleave within their recognition site, but at degenerate sequences; for example, *Mam*I cuts at -GATNN′NNATC-, where N can be any nucleotide. Nevertheless, both of these types of enzyme are still recognizably Type II on the basis of their biochemical properties. Except in the special cases just noted, all DNA molecules

Table 1.1. Characteristics of restriction and modification systems

	Class I	Class II	Class III
Composition	Multienzyme complex with R (endonuclease), M (methylase), and S (specificity) subunits, e.g. as R_2M_2S	Separate enzymes; endonuclease is a homodimer, methylase a monomer	M subunit provides specificity; on its own, functions as methylase; as heterodimer with R subunit, functions as methylase-endonuclease
Cofactors[a]	Mg^{2+}, ATP, SAM (needed for cleavage and methylation)	Mg^{2+}, SAM (for methylation only)	Mg^{2+}, ATP (for cleavage), SAM (needed for methylation; stimulates cleavage)
Recognition sites	Asymmetric, bipartite, may be degenerate, e.g. *Eco*K (AACN$_6$GTGC)	Symmetric, may be bipartite, may be degenerate (Table 1.2)	Asymmetric, uninterrupted, 5–6 nt long. E.g. *Eco*P15-CAGCAG. Two copies in opposite orientation, but not necessarily adjacent, needed for cleavage; one for methylation
Cleavage	Variable distance (100–1000 nt) from recognition site	Within recognition site, except for Class IIs (**s**hifted cleavage), which cleaves outside, at a defined distance	25–27 nt from recognition site
Number of systems characterized	Several, grouped into a few families. e.g. K, includes *Eco*B, *Eco*D, *Eco*K, and others	Hundreds	Few

[a] ATP: adenosine triphosphate; SAM: *S*-adenosyl methionine.

resulting from cutting with a given Type II enzyme will have the same sequences at their ends. That will not be true with the Types I and III enzymes, as they cut outside their recognition sites. Because molecules that are cut with Type II enzymes generally have the same ends, such molecules can base-pair with each other, and, as we shall see, be covalently joined by a DNA ligase. Some Type II enzymes give clean cuts rather than staggered ones, cutting both strands at the same place (see Table 1.2). This gives double-stranded or blunt ends on the molecules. That is not a problem, since blunt-ended

Table I.2.	Examples of recognition sequences of Type II restriction endonucleases[a]		
*Apa*I	G GGCC′C C′CCGG G	*Aha*III	T T T′A A A A A A′T T T
*Bam*HI	G′GATC C C CTAG′C	*Bgl*II	A′GATC T T CTAG′G
*Bsp*I20I	G′GGCC C C CCGG′C	*Dpn*I	GA′ TC CT′AG
*Dra*I	T T T′A A A A A A′T T T	*Eco*RI	G′A A T T C C T T A A′G
*Hinc*II	GTPy′PuAC CAPu′PyTG	*Hind*III	A′AGCT T T TCGA′A
*Hpa*II	C′CG G G GC′C	*Mae*III	′GTNAC CANTG′
*Not*I	GC′GGCC GC CG CCGG′CG	*Pvu*II	CAG′CTG GTC′GATC
*Sal*I	G′TCGA C C AGCT′G	*Sau*3A	′GATC CTAG′
*Sph*I	G CATG′C C′GTAC G	*Taq*I	T′CG A AGC′T
*Xba*I	T′CTAG A A GATC′T		

[a] ′=cleavage site; N=any nucleotide; Py and Pu=pyrimidine and purine nucleotides respectively.

molecules can also be joined by ligase. The following features of cleavage by Type II enzymes are also important:

1. **Recognition sites generally read the same on both strands** (as long as the same polarity, e.g. 5′ to 3′, is read). Such sequences are often described as palindromes. It is not necessary for recognition sequences to be palindromic for all molecules cut with the same enzyme to be able to reanneal, although it does increase the number of configurations in which reassociation can take place. For example, Figure 1.1 shows how two molecules cut with the enzyme *Hin*dIII (recognition sequence -AAGCTT-) can reanneal, with either end of the right-hand molecule annealing with the left-hand one. Two molecules cut with an enzyme with a non-palindromic recognition sequence could also reanneal, but fewer orientations are possible.
2. **Most enzymes have recognition sites of four or six nucleotides.** If all nucleotides occurred with equal frequencies (both in the DNA to be cut and in the enzyme recognition sites) and at random, a particular four-nucleotide motif would be expected to occur on average once every 4^4 (i.e. 256) nucleotides. So the average length of fragments generated by enzymes with such sites

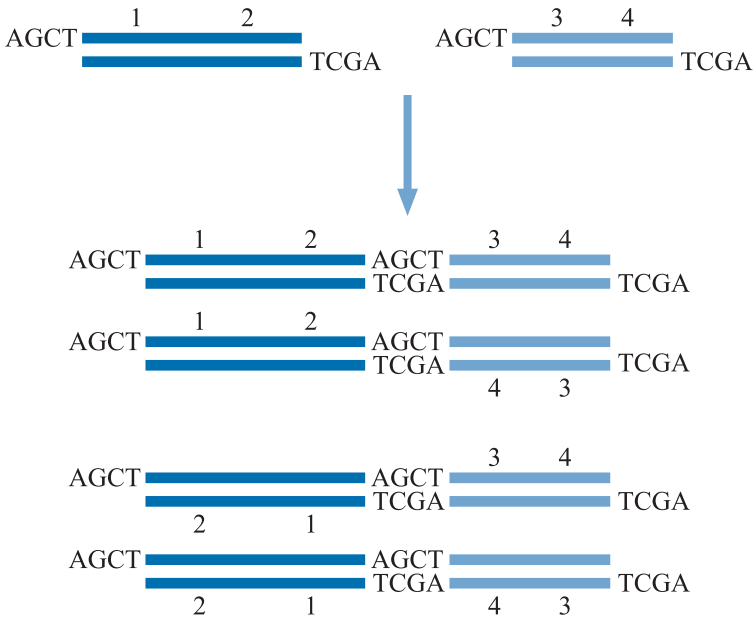


Fig I.1 Annealing of two molecules cut with *Hin*dIII. 1, 2, 3 and 4 represent arbitrary points on the molecules. Note that the right-hand molecule can anneal in two possible orientations to either end of the left-hand molecule, because of the palindromic nature of the *Hin*dIII cleavage site.

would be 256 base-pairs. Similarly, enzymes with a six-nucleotide recognition sequence would generate fragments with an average size of 4^6 (i.e. 4096) base-pairs. In practice, that does not happen for the following reasons:

(a) The bases do not occur with equal frequencies in the recognition sites.

(b) The bases do not occur with equal frequencies in the DNA to be cut, and the frequencies can vary over the genome.

(c) The bases do not occur at random, e.g. certain dinucleotides are favoured and others avoided. The degree of non-randomness often varies over a genome.

3. **Different enzymes can recognize the same sequence.** For example, *Dra*I and *Aha*III both recognize and cut at -TTT′AAA-. They are said to be isoschizomers. Enzymes with the same recognition sequence do not necessarily cut at the same position within it, though. For example *Apa*I recognizes and cuts at -GGGCC′C-, whereas *Bsp*120I recognizes and cuts at -G′GGCCC-.
4. **Different enzymes can generate the same ends**. For example, the enzyme *Sau*3AI produces the ends GATC-, and *Bam*HI does the same. This means that molecules produced by digestion with *Sau*3AI will be able to anneal and be ligated to molecules produced with *Bam*HI. Given that blunt-ended molecules can also be ligated, molecules cut with any enzymes that give blunt ends can be compatible. Notice, though, that ligation of molecules cut with different enzymes may not regenerate the original recognition sites used. It may also be possible to ligate (though at a lower efficiency) molecules whose sticky ends are nearly, but not fully complementary.

5. **Cutting can be influenced by other factors**. The most important are:
   (a) methylation;
   (b) the buffer used;
   (c) secondary structure in the substrate.

   Methylation of bases in DNA may result from the modification activity of a restriction–modification system, or from the activity of one of many independent methylases. Methylated bases commonly encountered include N^6-methyladenine, 5-methylcytosine, 5-hydroxymethylcytosine and N^4-methylcytosine. Restriction enzymes will generally not cut molecules where particular bases within their recognition site are methylated. Methylation at certain positions within the recognition site may not affect cleavage, and for some enzymes methylation at other positions may actually be required for cleavage. For example, cleavage by *Bam*HI is inhibited by methylation at the internal C of the -GGATCC- recognition site, but not by methylation of the other C or the A, whereas cleavage by *Apy*I is inhibited by methylation of the first C of its recognition site (-CCAGG- or -CCTGG-) and requires methylation of the second C.

   The specificity of some enzymes is affected by the buffer used. For example, the enzyme *Eco*RI normally cuts at the sequence -GAATTC-, but the specificity is relaxed in the presence of glycerol at concentrations greater than 5% v/v, and cutting can take place at -AATT- or -PuPuATPyPy-. This is often refered to as star activity, denoted *Eco*RI*. The extent of cutting can be modified by certain compounds. One example is ethidium bromide, shown in Figure 1.2. This molecule, which is also used for visualizing nucleic acids in gels, can be intercalated (inserted) between the bases in a double-stranded DNA molecule. This interferes with the action of restriction endonucleases, allowing cutting in one strand only.

   Some sites are cut much less efficiently than others within the same molecule. This may be due to secondary structures in the DNA that interfere with recognition or cleavage by the endonuclease.
6. **Enzyme activities are measured in units.** A unit is the amount of an enzyme required to digest a standard amount (usually 1 μg) of a standard type of DNA (often bacteriophage lambda, or a specified plasmid) in a given time (usually 1 h) under given

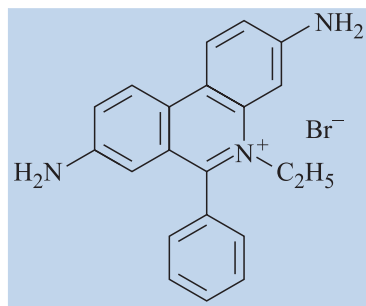


Fig I.2 Ethidium bromide.

conditions (temperature, pH, etc.). Digesting DNA molecules containing many sites may, therefore, require more units of enzyme than the amount required to digest the same mass of a DNA containing fewer sites.

7. **Restriction endonuclease preparations used for cloning must be free of other nucleases.** If not, the ends of the molecules generated might be degraded by exonucleases, and reannealing would be prevented. Contaminating endonuclease activity would cut the molecules into fragments with no (or the wrong) single-stranded ends, which would cause a similar problem. Manufacturers, therefore, usually test enzyme preparations by incubating DNA with a large excess of the enzyme, and determining what proportion of the products can be religated and whether the religated molecules can still be cut with the enzyme. The higher the proportion of correct religation, the 'cleaner' the enzyme preparation. Enzyme preparations can also be tested for the presence of exonuclease by incubation with DNA molecules that are radioactively labelled at their ends. The presence of exonuclease is indicated by the release of the radioactive label from the substrate DNA. Low levels of contaminating nucleases are not a problem in simple restriction enzyme mapping, though.
8. **Partial digestion may be useful.** Sometimes we deliberately do not carry out digestion with a restriction enzyme to completion. For example, we might need to fragment total DNA (often called genomic DNA) prepared from an organism into pieces of roughly the same size, say 10 kbp, so that every sequence in the organism is represented in the collection of 10 kbp fragments. Simply cutting to completion with an enzyme with a six-nucleotide recognition sequence and taking the fragments produced that were approximately 10 kbp would not be suitable. A lot of the DNA would only be cut into smaller (or larger) pieces and would never be represented in the 10 kbp size class. Therefore, a better method is to use an enzyme that cuts very frequently (e.g. at a four-nucleotide recognition site), but to adjust either the reaction time allowed or the ratio of enzyme to DNA in the reaction so that only a few of the possible sites are cut. In this way, the average size of the fragments can be raised to 10 kbp or whatever else is required; and if all sites have a more or less equal chance of being cut, then all regions of the DNA can be represented among the 10 kbp fragments (unless the distribution of sites in a particular region is grossly abnormal). This approach is very useful in constructing genomic libraries (see Chapter 5).
9. **Cleavage sites can be determined using standard molecules.** A selection of DNA molecules whose sequence is known completely are digested with the enzyme. Samples are also digested by combinations of the enzyme under test with others that have known cleavage sites. The sizes of the fragments generated are measured by gel electrophoresis, and a computer

analysis allows you to infer possible recognition sites from the sequences on the grounds that they are the only sites that would generate fragments of the observed sizes. More accurate measurement (to the exact number of nucleotides) of the sizes of molecules generated by digestion allows the actual cleavage site within the recognition sequence to be inferred. These accurate size measurements are also done electrophoretically, using the products of DNA sequencing reactions as size markers.

10. **Nomenclature follows a simple convention.** Once an enzyme has been characterized, it must be given a name. The convention is that names start with three letters (italicized): the first letter of the genus and the first two letters of the species of the source cells. Where relevant, they are followed by an indication of the strain and then a number (in Roman numerals) indicating which one of the enzymes from that strain the name refers to. For example, the enzymes *Eco*RI and *Eco*RII refer to the first and second activities isolated from strain R of *E. coli*. Often, names are abbreviated, so *Eco*RI is often referred to colloquially just as 'RI' (pronounced 'R-one'), *Bam*HI as 'Bam' and so on. According to the general conventions for enzyme nomenclature, Types I, II and III restriction endonucleases are classified as 'endodeoxyribonucleases producing 5′-phosphomonoesters' and classified as EC 3.1.21.3, EC 3.1.21.4 and EC 3.1.21.5, respectively.
11. **Other enzymes can cut DNA molecules at specific sequences.** During the course of infection of *E. coli* cells by bacteriophage lambda, copies of the phage genome are cut at a specific 16-nucleotide site, called *cos*, leaving a 12-nucleotide single-stranded overhang. The cleavage is carried out by a phage enzyme called the terminase. *Cos* sites are sometimes introduced into large DNA molecules to allow cleavage at a single site when conventional restriction enzymes would cut the molecule into several pieces.

1.2.2 DNase

In some instances, restriction endonucleases are unsuitable for cutting DNA. That might be so if the DNA has a very abnormal base composition, although such a wide variety of enzymes is now available, with so many recognition sites, that this is rarely a problem. A more common problem is when it is necessary to break DNA into a random collection of fragments with a mean size of only a few hundred base-pairs. Partial digestion with a four-nucleotide-recognizing enzyme is not suitable; nearly every site would have to be cut to get the required average size, and this would mean that some sequences would be represented only on fragments either much smaller or much larger than the required size range. The problem can be avoided using a DNase such as DNaseI, which has very little (and for this purpose essentially no) sequence specificity. Again, careful adjustment of either the enzyme:DNA ratio or the incubation

time is necessary to ensure the optimal distribution of fragment sizes. One problem with the use of DNase is that the ends of the molecules produced do not have a unique single-stranded sequence. Also, not all the ends are blunt. This makes cloning of the fragments difficult, but the problem can be solved by rendering all the ends blunt with a suitable DNA polymerase.

1.2.3 Physical stress

In addition to enzymatic means, we can use physical shearing to cleave DNA at random. We can accomplish this in several ways. For example, we can simply stir a solution or force it through a narrow opening such as a syringe needle or a pipette tip, or we can use sonication (which provides high-frequency vibrations). In practice, sonication is the preferred method, since it is the easiest to control and is often more reproducible than DNaseI treatment. Different kinds of sonicator are available. In the simplest form, a metal probe is dipped into the solution and vibrates at high frequency. This has the disadvantage that the probe can be a cause of cross-contamination between DNA preparations unless it is carefully cleaned. An alternative instrument is the cup-horn sonicator, where the solution to be sonicated is retained in a tube that floats in a small volume of water. The probe is dipped into the surrounding water and vibrations are transmitted through the water to the tube containing the sample. With shearing, as with DNase treatment, there is no control over the sequences at the ends of the fragments produced.

1.3 Modification

1.3.1 Phosphatases

Phosphatases are enzymes that hydrolytically remove phosphate groups from DNA molecules, replacing them with hydroxyl groups. The terminal phosphate groups left by restriction enzymes are needed for most ligation reactions, and the application of phosphatase in blocking unwanted ligation reaction will be described in Chapter 3. Widely used preparations come from calf intestines, shrimps and the Antarctic psychrophilic (cold-loving) bacterium TAB5. Many phosphatase preparations, especially the last one, can be readily inactivated by heating. This is useful when we want to terminate phosphatase activity prior to a ligation.

1.3.2 Polymerases

We will meet four classes of DNA or RNA polymerases: DNA-dependent DNA polymerases, RNA-dependent DNA polymerases, DNA-dependent RNA polymerases and template-independent polymerases. There are also RNA-dependent RNA polymerases, but they are less important for our purposes.

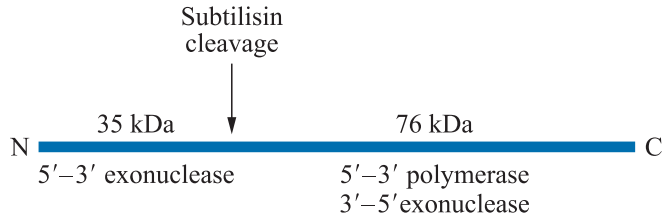


Fig I.3 DNA polymerase I. The locations of the activities and the cleavage site of subtilisin are indicated.

1. **DNA-dependent DNA polymerases.** These enzymes synthesize a DNA strand in a 5′-3′ direction using a DNA template. They can also have 5′-3′ and 3′-5′ exonuclease activities, and all these activities can be exploited in various ways. The preparations used come from bacteria, such as *E. coli* and *Thermus aquaticus*, and from bacteria infected with viruses such as T4 and T7. The *E. coli* enzyme that is widely used is DNA polymerase I, which normally has important roles in DNA repair and the replacement with DNA of the RNA primers used for DNA synthesis. The enzyme has 5′-3′ polymerase, 3′-5′ exonuclease (serving a proof-reading function), and 5′-3′ exonuclease activities. These are essentially located on different domains of the molecule. Cleavage with the protease subtilisin generates an N-terminal fragment of 35 kDa containing the 5′-3′ exonuclease activity and a C-terminal one of 76 kDa with the polymerase and 3′-5′ exonuclease activities (Figure 1.3). The 76 kDa piece is sometimes called the Klenow fragment and the intact molecule the Kornberg enzyme.

The 5′-3′ DNA polymerase activity allows a complementary DNA strand to be synthesized using a suitable template. This template might be a large piece of single-stranded DNA with a small primer annealed, or it might be a restriction fragment with a recessed 3′ (i.e. overhanging 5′) end. Incubation of either of these templates with DNA polymerase and the correct deoxynucleoside triphosphates would result in the filling in of the single-stranded region (the 'recessed end') to produce a blunt-ended molecule. This is often called end filling. Note that the 5′-3′ exonucleolytic activity of the Kornberg enzyme could also produce such a molecule by degradation of the overhanging 5′ end rather than by synthesis of its complement. A recessed 5′ end cannot be rendered blunt by the polymerase activity, because synthesis would have to be in the 3′-5′ direction, which is not possible. In that case, the 3′-5′ exonucleolytic activity could render the ends blunt by degradation of the overhanging 3′ end. Rendering overhanging ends blunt by any of these means is termed polishing, and is summarized in Figure 1.4. *E. coli* DNA polymerase is also used for DNA sequencing.

Thermostable DNA polymerases are particularly important for amplification of DNA by the polymerase chain reaction (PCR), as described in Chapter 2. They are isolated from extremely thermophilic bacteria, often growing in hyperthermal oceanic vents, such as *Thermus aquaticus*, *Thermococcus litoralis* and *Pyrococcus furiosus*.

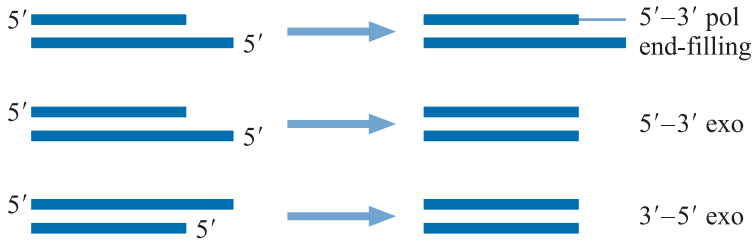


Fig 1.4 Polishing overhanging ends using polymerase (pol) or exonuclease (exo) activity. The thin line indicates DNA newly synthesized by polymerase activity.

Some of these bacteria grow at temperatures of over 100°C, and their DNA polymerases can also function effectively in vitro at very high temperatures. The first enzyme commonly used in PCR was from *Thermus aquaticus*, with a temperature optimum of 75–80°C. More details of the enzymes used for PCR are given in Chapter 2. Thermostable polymerases can have other applications, such as in DNA sequencing.

The viral DNA polymerases are encoded in the viral genomes and are synthesized upon infection of *E. coli* cells. The polymerases are essential for replication of viral DNA in the absence of host DNA synthesis. T4 DNA polymerase has a particularly active 3′-5′ exonuclease function (more than 200 times that of the Klenow enzyme) and is especially useful in polishing up sticky ends. T7 DNA polymerase is sometimes used in DNA sequencing, where it generates fewer artefacts than does the Klenow fragment of *E. coli* DNA polymerase, probably as a result of higher processivity (i.e. a lower tendency to dissociate from the template). It has no 5′-3′ exonuclease activity. The preparation used for sequencing usually has the 3′-5′ exonuclease activity inactivated. This was first done chemically, but subsequently by manipulation of the polymerase gene, to produce a form of the enzyme called 'Sequenase'.

2. **RNA-dependent DNA polymerases.** These enzymes are also known as reverse transcriptases, often abbreviated to RTases, owing to their ability to reverse the usual flow of information by transcription in the 'central dogma'. They are encoded by retroviruses (in the *pol* gene), which have an RNA genome that has to be turned into DNA as part of their life cycle. (The DNA is subsequently inserted into the host's genome.) Enzymes from avian myeloblastosis virus (AMV) and murine Moloney leukaemia virus (M-MLV) are widely used. Modified versions are also available, for example, with enhanced thermostability. Like other DNA polymerases, RTases need a primer that is hydrogen bonded to the template, and they direct DNA synthesis in a 5′-3′ direction. In virally infected cells, the primer is a cellular tRNA molecule that shows complementarity to part of the retroviral genome. In the laboratory, a suitable primer must be supplied. The most important use of RTase is in the synthesis of DNA from RNA when making cDNA, for example, as described in Chapter 5. RTases can also be used in sequencing,

either by direct sequencing of RNA, provided a suitable primer is available, or (more rarely) in DNA sequencing, since they can use DNA as well as RNA as a template. RTases do not have proof-reading 3′-5′ exonuclease activity and, therefore, have a higher error rate than most DNA polymerases. Under standard conditions this can be as high as 1 nucleotide in 500 being incorrect, and this should be taken into account when using clones obtained with reverse transcriptase (although it often is not).

3. **DNA-dependent RNA polymerases**. As well as encoding DNA polymerases, viruses can encode their own RNA polymerases. Because the phage RNA polymerases are very specific for phage promoters, viruses can inactivate the host RNA polymerase while maintaining transcription of their own genomes. This means that more ribonucleotides, amino acids and so forth are available for synthesis of viral proteins. The phage RNA polymerases, such as those from T3, T7 and SP6, are used in conjunction with the appropriate phage promoters to direct high levels of very specific transcription. They can be used in vivo or in vitro.
4. **Template-independent polymerases**. Some enzymes can add one or more nucleotides to a molecule without depending on a template. Two enzymes are particularly important. One is terminal transferase, from calf thymus. The enzyme can attach a series of deoxyribonucleotides one by one to the 3′ end of a DNA molecule. This is used to add tails of a single nucleotide to existing DNA molecules. Another example of a template-independent polymerase is the *Taq* polymerase used in the PCR. This adds a single dA-residue to the end of a PCR product, and this is exploited during cloning of PCR products.

1.3.3 Exonucleases

We have already dealt with *endo*nucleases, and their role in cutting DNA up into discrete fragments, and also with the use of the *exo*nucleolytic functions of DNA polymerases. Cells also contain exonucleases that do not have any associated polymerase activity, but nevertheless have a wide range of roles. We can exploit them for the removal of single-stranded ends, which we have already discussed, and also for the shortening of double-stranded DNA. This can be used in constructing collections of clones for DNA sequencing and for targeted mutagenesis (see Chapter 7). Exonucleolytic shortening of double-stranded DNA can be carried out in two ways.

1. **Removal of strands separately.** One technique for shortening double-stranded DNA involves removing one strand at a time. The DNA is cut with a restriction enzyme that leaves protruding 5′ ends. Then it is treated with two separate enzymes. The first is commonly exonuclease III (exoIII). This enzyme has 3′-5′ exonuclease activity on double-stranded DNA. Therefore, it will degrade the recessed 3′ ends of the molecule described, leaving

a single-stranded region at both ends. The DNA molecules are then treated with a single-strand-specific exonuclease (i.e. one that will degrade single-stranded DNA only), such as S1 nuclease from *Aspergillus oryzae* or mung bean nuclease. This polishes up the single-stranded ends left by exoIII.

As described, this method leads to a molecule being shortened from both ends, and this is not always desirable. It is possible to protect one end of a molecule from degradation by exploiting the fact that exoIII will not attack *protruding* 3′ ends (as these are effectively just single-stranded molecules and the enzyme works only on double-stranded molecules). The target molecule is cut with two restriction enzymes chosen so that the resulting fragment to be shortened has one protruding 3′ end and one recessed end. Subsequent treatment with exoIII degrades only the latter end, which is then polished with S1 or mung bean nuclease (Figure 1.5).

Protection from exoIII can also be achieved by replacing one nucleotide at one end by a phosphorothioate nucleotide analogue, which is resistant to removal by exoIII. (Incorporation of the analogue can be achieved by digestion with an enzyme followed by end-filling in the presence of the analogue.)

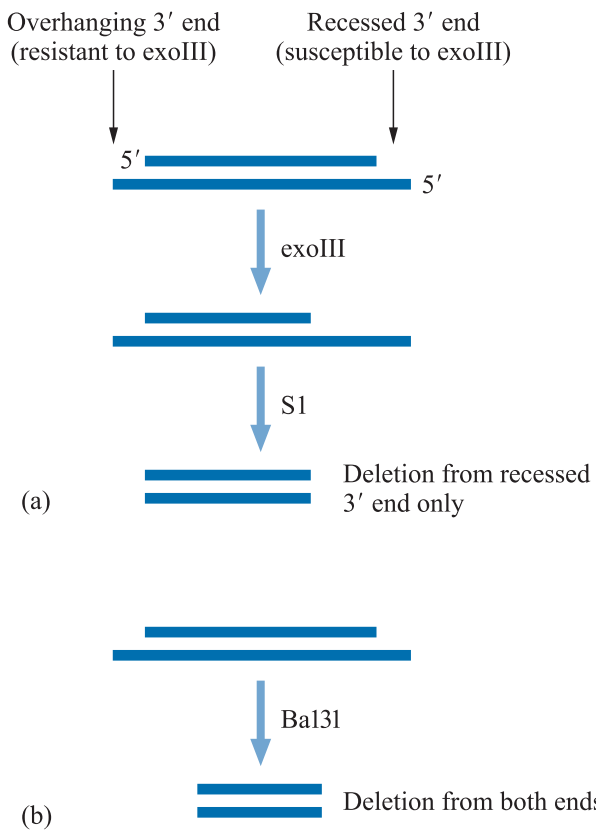


Fig 1.5 Exonucleolytic shortening of dsDNA. (a) Use of exoIII and SI nuclease leads to deletion from recessed 3′ ends (overhanging 5′ ends) but not overhanging 3′ ends (recessed 5′ ends). (b) Use of Bal3I leads to deletion from both.

2. **Removal of both strands together**. A second approach for shortening double-stranded DNA molecules is to remove both strands of the duplex concurrently (Figure 1.5). One exonuclease that can do this is Bal31. Its main activity is a 3′-5′ exonuclease, which generates single-stranded ends. These are then degraded by an endonuclease activity that the enzyme also has. Treatment of DNA with Bal31 will, therefore, lead to simultaneous removal of nucleotides from both strands at both ends. In practice, some overhangs may be left, as the endonuclease lags behind the exonuclease.

In both methods for exonucleolytic shortening, the extent of the deletion will depend on the amounts of nuclease added and the incubation time. Controlling these parameters, therefore, allows the extent of the deletions to be manipulated. By taking samples from the reaction after different times, one can produce a 'nested' series of deletions of a range of lengths. If a single deletion is required, then it may be simpler to use chemical synthesis to produce the required molecule.

1.3.4 Methylases

We have already seen that the modification side of restriction–modification systems protects a cell's DNA by methylation at the recognition site. Sometimes in cloning it is also necessary to protect DNA against cleavage by a particular enzyme. This can be done by treating the DNA with an appropriate methylase, which transfers methyl groups onto the DNA from *S*-adenosyl methionine. For example, protection against digestion by *Eco*RI could be conferred by treatment with *Eco*RI methylase.

1.4 Ligation

Ligation is the alignment of the ends of two (usually double-stranded) DNA molecules and the formation of a covalent linkage (phosphodiester bond) between them in one or both strands. A break in the sugar–phosphate backbone of a double-stranded DNA molecule that can be sealed simply by the formation of a phosphodiester bond is called a nick. If nucleotides are missing, then it is called a gap and cannot be sealed by ligation alone.

1.4.1 Categories of reaction

Ligation reactions may be blunt ended or sticky ended. In the former, the molecules to be joined do not have overhanging single-stranded ends, which would have the potential to reanneal. The ends might, for example, have been generated directly by the action of a restriction endonuclease that gives a straight cut, or by polishing the ends of molecules produced with an enzyme that generates staggered cuts. In sticky-ended ligation, the molecules have

complementary single-stranded ends. The ends can base-pair, and ligation then forms the phosphodiester bond(s) to seal the nicks. The reaction is most efficient if the sticky ends complement each other exactly. A small amount of mismatch may be acceptable, though. In general, ligation of correctly matched ends is more efficient than ligation of blunt ends. However, the energy required to break the few hydrogen bonds holding sticky ends together is very low, being comparable to the vibrational and kinetic energy of molecules at room temperature. Therefore, sticky-ended ligations are traditionally done at lower than room temperature, usually at 4°C.

Ligation reactions may be intermolecular, in which the end of one molecule is ligated to the end of another, or intramolecular, in which the end of one molecule is ligated to the other end of the same molecule, resulting in its circularization. The former requires collision between two separate molecules and implies second-order kinetics; the latter implies first-order kinetics. Increasing the concentration of the components of a reaction will increase the probability that two separate molecules will collide, but will not alter the probability that one end of a molecule will meet the other one. Increasing the concentration should, therefore, enhance the frequency of intermolecular ligations compared with intramolecular ligations. The effective concentration can be increased by adding volume excludants, such as polyethylene glycol.

1.4.2 Enzymes for ligation

Ligases are the most commonly used enzymes for carrying out ligations. They are part of the routine battery of enzymes required by a cell for the maintenance of its DNA. They are used in joining together adjacent Okazaki fragments produced in replication, and in sealing the nicks that arise from damage and repair processes. Like the DNA polymerases we looked at earlier, the ligases that we will use can come from normal *E. coli* or from cells that have been infected by viruses.

1. **T4 DNA ligase** is encoded by bacteriophage T4, and is produced on infection of *E. coli* cells. It can carry out both blunt-ended and sticky-ended ligations, and requires ATP. It requires a 3′-hydroxyl and a 5′-phosphate group on the molecules to be joined.
2. ***E. coli* DNA ligase** is the endogenous bacterial enzyme. Unlike T4 DNA ligase, it is unable to carry out blunt-ended ligations (or does so only very inefficiently) and, therefore, is particularly useful if such ligations need to be avoided. This might be the case if you were trying to seal nicks in damaged DNA without also joining non-contiguous sequences. Like the T4 ligase, *E. coli* DNA ligase requires a 3′-hydroxyl and a 5′-phosphate group, but it requires nicotinamide adenine dinucleotide (NAD^+) as a cofactor. Essentially the same reaction is catalysed by each of these ligases; this is shown in Figure 1.6. In both cases, adenosine monophosphate (AMP) is added to the 5′-phosphate, of one DNA molecule

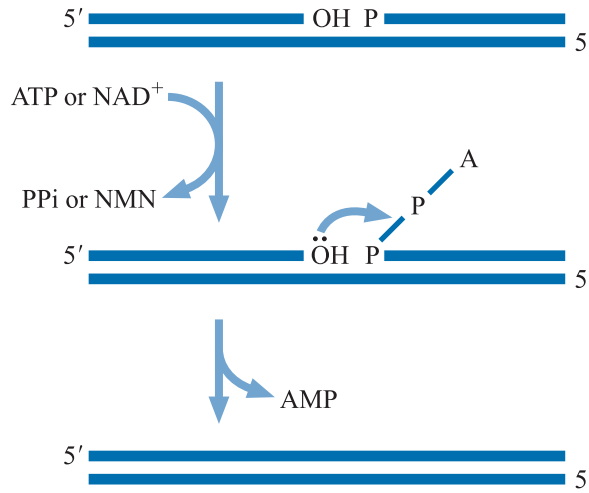


Fig I.6 Mechanism of ligation by T4 DNA ligase (using ATP) or *E. coli* ligase (using NAD^+).

liberating either pyrophosphate from ATP or nicotinamide mononucleotide from NAD^+. The AMP is then displaced in a nucleophilic attack by the 3′-hydroxyl of the other DNA molecule.

3. **Topoisomerase** is another enzyme with DNA ligase activity. The normal function of these enzymes is to alter the degree of supercoiling of DNA molecules. They do this by cleaving one or both strands, rotating the duplex, and resealing it. Given a linear DNA molecule with topoisomerase attached to the end and a suitable target molecule, the enzyme will ligate the two. This allows ligation to be accomplished more rapidly than with conventional DNA ligase.
4. **Transposase.** Transposable genetic elements are able to move from one piece of DNA to another, under the action of a transposase enzyme. This can be used as a way of inserting features such as origins of replication or antibiotic-resistance genes into a molecule, but it is rather specialized.
5. **Recombinase.** There are a number of phage-based recombination systems that catalyse breakage and rejoining of molecules at specific sites. For example, bacteriophage lambda has such a system to direct recombination between the phage genome and the chromosome of an infected bacterium to catalyse the insertion of the former into the latter. The integration takes place by recombination between a site on the phage genome (*attP*) and a site on the bacterial genome (*attB*) to generate a phage–bacterial DNA chimera. The *cre–lox* recombinase system of bacteriophage P1 is frequently used for the deletion of regions of DNA flanked by *loxP* sequences, and has applications in the modification of gene expression in transgenic animals.

That concludes our survey of the enzymes that we will need to use in the forthcoming chapters. There are a few techniques that need to be mentioned before we can really start, because they are particularly relevant. One of the most fundamental for any cloning project is the

introduction of DNA into a recipient organism. This is called transformation. A wide range of methods are available for this, and many methods are applicable to several different groups of organisms, so it is reasonable to consider transformation at this stage. We will look briefly at the purification of plasmids from *E. coli*, since this is another of the most frequent operations in a cloning project. We will also cover gel electrophoresis of nucleic acids, since many general textbooks do not cover all aspects adequately for our purposes, and blotting. We will discuss DNA sequencing and look briefly at the synthesis of oligonucleotides with defined sequences. We will not go into details of the purification of DNA, RNA or protein from cells other than *E. coli*, since the protocols vary widely from one organism to another, and long lists of practical instructions would not be appropriate here. Nor will we consider general molecular biological techniques, such as transcript mapping and gel retardation assays. Although they can be important in the analysis of cloned genes, they are now quite routine throughout molecular biology and more general textbooks give details.

1.5 Transformation

Cloning projects almost inevitably require the introduction of DNA molecules that have been generated in vitro, using endonucleases, ligases and so forth, into a recipient organism, termed a host. Many methods are available, and the process is commonly called transformation. Other terms, such as transfection, are also used. Transfection often refers to uptake of viral DNA (which may sometimes be packaged in a viral coat for this process) and also to the transformation of eukaryotic cells. We will discuss the general principles here and give more specific details where appropriate in later chapters.

1.5.1 'Natural' methods

Many bacterial species have a natural ability to take up exogenous DNA molecules. This ability is sometimes called competence and may be limited to a particular growth phase of a bacterial culture, and be associated with the induction of a specific set of bacterial proteins. It seems likely that competence may have been developed to allow bacteria to acquire new biochemical abilities under conditions of nutrient deprivation (or other stresses), with the possibility that these abilities may include metabolism of other available nutrients. If we are dealing with a bacterial species that is naturally transformable, then it may be possible, therefore, to get cells to take up DNA in the laboratory simply by incubating them (perhaps at a specific growth phase) with the DNA.

Although many species do not take up naked DNA naturally, they may be able to acquire DNA from other cells by conjugation.

Conjugation is a process by which plasmids can be passed from one cell to another through direct physical contact, and the machinery for doing this is often encoded by the plasmid being transferred. A good example of this is the F (fertility) factor of *E. coli*. It can be transmitted from one cell to another through the sex pilus, which is encoded by the F plasmid. In many cases, it is not necessary for the bacteria involved in conjugation to belong to the same species. Some plasmids can be maintained in a wide range of bacterial hosts and be passed from one to another. These are called broad host range plasmids. We may, therefore, be able to get DNA into a bacterial species (A) that is not readily transformable by using a plasmid to transform a species that is transformable (B) and relying on conjugation for the transfer of the plasmid from B to A. A plasmid that contains the genes necessary to direct its own transfer into another cell is said to be self-mobilizable. Some plasmids do not contain all the information for their own transfer, but can nevertheless be transferred if the genes are present elsewhere in the host. Such plasmids are mobilizable.

One particularly important transformation procedure relies on the transfer of DNA from a bacterium into a eukaryotic host cell. This is the use of the plant pathogen *Agrobacterium tumefaciens* to insert DNA into plant cells. This is discussed in more detail in the context of plant genetic manipulation in Chapter 9.

Viral infection is a third process by which cells can take up DNA. The infection process usually relies simply on the viral coat and is independent of the nucleic acid contained within it. If we can arrange for the DNA that we wish to be introduced into a host cell to be packaged into the coat of a virus that infects it, then we can get very efficient insertion of the DNA into the target cell; this is more efficient (measured as the fraction of DNA molecules supplied that successfully enter the host) than transformation. The use of DNA in a viral coat in this way is sometimes called transfection, although the term can also refer to the introduction of molecules based on viral DNA by direct transformation.

1.5.2 Chemical transformation

The efficiency of uptake of DNA can be increased by a range of chemical treatments. For example, under normal conditions, *E. coli* is not readily transformable, but treatment with ice-cold calcium chloride solutions followed by a heat shock induces DNA uptake. The reasons for this are not clear. Treatment with more complex mixtures of chemicals, containing manganese, hexamine cobalt and rubidium or potassium ions as well as calcium ions, together with dimethyl sulphoxide and dithiothreitol is also used. This can bring about a higher efficiency of DNA uptake, but electroporation (see below) may cause an even higher efficiency of uptake. Competent *E. coli* cells can often be deep-frozen for future use with relatively little loss of viability or competence (and they are available

commercially), which can greatly reduce the amount of labour involved in preparing them. For some bacteria, such as the Gram positive *Bacillus* and *Streptomyces* species, it may be helpful to start by removing the cell wall enyzmatically to generate protoplasts prior to adding the DNA. However, it will then be necessary to allow cells to regenerate their walls. Wall regeneration can often be achieved by plating the cells on a suitable growth medium.

A variety of chemical treatments is used for inducing eukaryotic cells to take up DNA, as described in Chapter 9. One approach is to associate the DNA with lipids to form vesicles. These vesicles will then fuse with the cell membrane, delivering the DNA into the cell. Lipids with spermine headgroups are sometimes used to improve the interaction with DNA. Mixing the DNA with modified polysaccharides, such as diethylaminoethyl (DEAE)-dextran is also used for chemical transformation, as is co-precipitation of the DNA onto the surface of cells with calcium phosphate.

1.5.3 Electroporation

Many cells, both prokaryotic and eukaryotic, can be induced to take up DNA by subjecting them to an electric shock. This is termed electroporation. Cells are generally placed in a cuvette between two electrodes and a potential difference (i.e. a shock) applied across the electrodes. The potential difference is usually generated by charging a capacitor and then allowing it to discharge across the electrodes. The parameters to be altered are the field strength of the shock (the maximum potential difference divided by the separation between the electrodes) and the rate of decay of the field. This will be exponential and is determined by the product of the resistance of the total circuit and the capacitance. It should be varied according to the particular cells used. It seems likely that the shock disturbs the structure of the membrane, making it briefly permeable. The treatment usually leads to death of a fraction of the cells. Establishing optimal conditions for electroporation usually requires a balance between cell death and the fraction of cells that are induced to take up DNA, both of which increase as the severity of the conditions increases.

1.5.4 Physical methods

There are also physical methods for getting DNA into recipient cells. One such method uses a particle gun that literally fires DNA into recipient cells. The DNA is adsorbed onto tiny microprojectile beads, often of gold or tungsten, a micrometre or so in diameter. An explosion in the gun propels a macroprojectile forward, which in turn propels the microprojectiles. The macroprojectile is stopped by a perforated plate, but the microprojectiles pass through the holes and enter the target tissue, which is placed beyond the plate. The explosion in the gun can be generated either by a gunpowder charge or by compressed gas. Transformation using the particle gun

is also called biolistic transformation. Another physical transformation method is to vortex cells with silicon carbide whiskers or glass beads and the target DNA. Presumably the whiskers and beads form tiny transient holes in the cell membrane through which DNA can pass. Microinjection of DNA through a tiny syringe (a femtosyringe) with a tip diameter of 0.1 μm has also been used.

1.6 Purification of plasmid DNA from *E. coli*

Probably one of the most frequent operations in cloning work is the preparation of plasmid DNA from *E. coli*. Isolation of genomic DNA from other organisms follows similar principles. There are a number of different techniques, but essentially they involve the same operations:

(a) cell lysis
(b) removal of proteins and chromosomal DNA
(c) collection of plasmid DNA
(d) further purification if necessary.

There are many laboratory manuals that provide practical details, so we will settle for a brief outline. Commonly, cells are lysed by treatment with lysozyme (hydrolysing the cell wall, which comprises chains of *N*-acetyl glucosamine and *N*-acetyl muramic acid crosslinked by short peptides) and detergents. The 'chromosomal' DNA (the bacterial genomic DNA) is removed by centrifugation (as the molecule is so large it can be pelleted by centrifugation in a benchtop microcentrifuge) and the proteins are removed by extraction into phenol and chloroform. (Variations on this include the use of boiling, sodium hydroxide and sodium dodecyl sulphate to denature the proteins and chromosomal DNA in order to facilitate their removal by centrifugation.) The nucleic acid that remains in solution is then precipitated, usually by the addition of sodium or ammonium acetate and ethanol. The nucleic acid precipitated in this stage includes plasmid DNA, any chromosomal DNA that centrifugation did not remove, and RNA. Chromosomal DNA contamination should be low, but if necessary it can be removed by treatment with an exonuclease, because it is likely to be present as sheared linear fragments. Circular plasmid DNA will be resistant to exonuclease digestion, as it lacks free ends. Contaminating RNA can be destroyed if necessary with ribonuclease. These methods will leave us with plasmid DNA preparations that are usually suitable for use without further purification.

DNA may also be separated from other cellular material by adsorption onto a suitable solid-phase support (usually in a column) followed by elution. This method avoids the need for phenol purification, and may also remove RNA. One approach uses a column containing silica that binds DNA in the presence

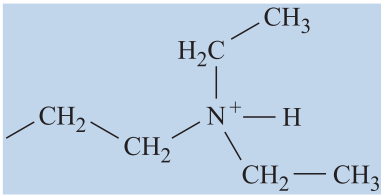


Fig 1.7 Structure of the diethylaminoethyl group.

of a high salt concentration and chaotropic agents (which help less polar molecules to remain in the aqueous phase). The DNA is retained by the column but under these conditions the polysaccharides and proteins pass through. After washing the column, the DNA can be eluted using a buffer with low salt concentration. An alternative approach to DNA purification uses an ion-exchange resin based on silica derivatized with DEAE groups (Figure 1.7). These groups are positively charged and bind DNA and other molecules, such as RNA and proteins. Different molecules are eluted from the ion-exchange column by varying the ionic conditions, allowing purification of DNA away from the contaminants.

If further purification of DNA is necessary, whether prepared by phenol purification or column chromatography, caesium chloride density gradient centrifugation can be used. This relies on the fact that a solution of caesium chloride, if spun in an ultracentrifuge, will form a concentration (and density) gradient because the caesium ions are so heavy. The nucleic acids and caesium chloride are dissolved and subjected to centrifugation. The gradient of caesium chloride concentration and density then forms, and the DNA migrates to the position in the gradient where its own density (determined by whether the DNA is supercoiled and by the G+C content, with a higher G+C content giving a greater density) is equal to that of the surrounding solution. This is an equilibrium position, and further centrifugation will not alter the position of the DNA. The closed circular plasmid DNA forms a band at one position in the centrifuge tube. Nicked (and, therefore, not supercoiled) plasmid DNA and linear fragments of chromosomal DNA form a band higher in the tube. RNA forms a pellet at the bottom of the tube and any protein left floats to the surface. The location of the plasmid DNA is determined by including the orange–red stain ethidium bromide in the solution before centrifugation. The ethidium bromide molecule is very flat and hydrophobic (Figure 1.2), and intercalates between the stacked bases of a piece of double-stranded nucleic acid. Once inserted, it will fluoresce very brightly with a rosy pink colour under ultraviolet (UV) light, because energy is transmitted to it from the DNA itself. The fluorescence can be used to locate the DNA in the centrifuge tube. However, there is often so much ethidium bromide intercalated in the plasmid DNA that UV illumination is not necessary. Once the DNA has been collected, the ethidium bromide is removed by washing the solution with butanol. Caesium chloride is removed by dialysis and the DNA precipitated and redissolved.

1.7 Gel electrophoresis of nucleic acids

1.7.1 The principle

The basic use we will make of gel electrophoresis is to separate nucleic acid molecules by size (although the technique is equally applicable to proteins). In essence, a gel that contains buffer is formed by a meshwork of molecules, and nucleic acids are driven through it by an electric field. At the pHs used, the charge on a DNA molecule results from the negatively charged phosphate groups of the sugar–phosphate backbone. So the charge on the molecule and, therefore, the force attracting it to the positive terminal will be directly proportional to the number of phosphate groups on it and, therefore, to its length. The mass of the molecule will also be proportional to its length (if we ignore the effects caused by the differences in molecular weights of the bases, which are negligible). The force per unit mass on a molecule and, therefore, the acceleration towards the anode will, consequently, be independent of the size of the molecule. If unhindered, all molecules would, therefore, move in the electric field at the same speed. However, the gel matrix retards the progress of the molecules, which in effect become entangled in it. Small linear molecules are able to pass through more freely than large ones, and so move faster through the gel. Therefore, the DNA molecules are separated by size, with smaller ones moving further. Electrophoresis is carried out for a given length of time, and the positions of the DNA visualized (generally by staining the gel with ethidium bromide or by autoradiography if the DNA is radioactive). Through comparison of the positions with those of standards of known sizes run in the same gel, the sizes of molecules under analysis (e.g. the products of a restriction digest) can be estimated. It is found empirically that, for linear molecules, a graph of the distance moved against the logarithm of the size of the fragment gives a straight line over a reasonable range. The behaviour of circular molecules is not as simple as that of linear molecules, and the retardation by the gel matrix depends on the topological parameters of the molecule (the degree of supercoiling, etc.). Different topological forms can be distinguished by their mobilities, but absolute size estimates are generally unreliable.

1.7.2 Gel matrices

Two gel matrices are favoured: agarose and polyacrylamide. Agarose is a purified polysaccharide from rhodophyte ('red') algae of the genera *Gelidium* and *Gracilaria*, and is composed of modified galactose residues. (In conjunction with another polysaccharide, agaropectin, it forms agar.) The agarose is dissolved by boiling in buffer and then allowed to cool. This results in the formation of a gel at room temperature. Polyacrylamide is formed by the polymerization of acrylamide and *N*,*N*′-methylene-bis-acrylamide, which

crosslinks the chains. The structures of these molecules are shown in Figure 1.8. Polymerization is brought about by the addition of ammonium persulphate and tetramethylethylenediamine (TEMED). The TEMED catalyses the formation of free radicals from the persulphate ions, and these initiate polymerization of the acrylamide. Agarose gels are usually cast and run in flat trays, and polyacrylamide gels are usually cast and run vertically between two glass plates. Because of the effective pore sizes in the gel matrices and the concentrations of polyacrylamide or agarose that it is convenient to work with, polyacrylamide gels are best suited for electrophoresis of molecules up to 1 kbp, and agarose gels are best suited for molecules from a few hundred nucleotides and up. Of course, some variation in the concentrations used is possible to suit the requirements of the experiment.

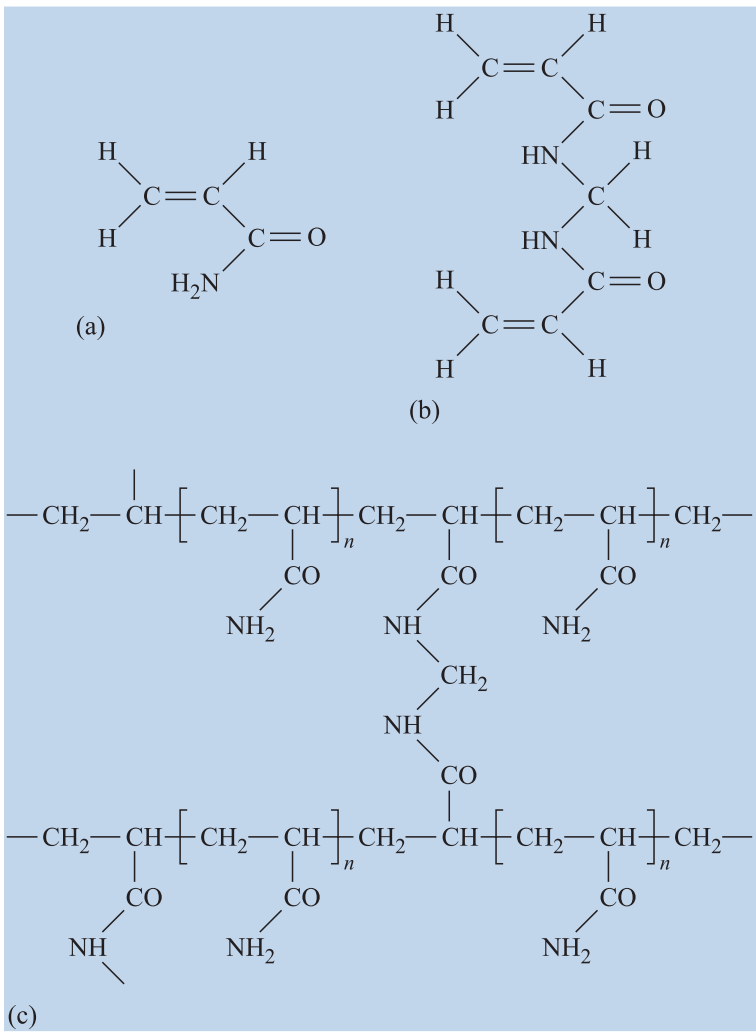


Fig 1.8 Structures of acrylamide (a), *N*,*N*′-methylene-bis-acrylamide (b), and polyacrylamide (c).

For electrophoresis of *single-stranded* DNA or RNA, denaturing gels are usually necessary in order to stop complementary strands from reannealing and intramolecular base-pairing from forming secondary structures, which would affect the mobility of the nucleic acids. Denaturing electrophoresis is particularly important in DNA sequencing, in which the ability to resolve molecules differing in length by a single nucleotide is necessary. (However, automated DNA sequencing systems usually employ capillary electrophoresis rather than conventional gel electrophoresis.) Denaturing gels can be run using exactly the same matrices that are used with nondenaturing ones, with the addition of denaturing agents. These interfere with the formation of secondary structure in the nucleic acids, and include urea, formamide and formaldehyde. Electrophoresis at elevated temperature (the heating caused by the electric current in the gel may be sufficient) also helps to disrupt secondary structure.

1.7.3 Recovering DNA from gels

It is often useful to be able to recover material, usually DNA, from gels. Perhaps one needs to clone a specific restriction fragment from a digest by separating the desired fragment from others using a gel and then recovering the fragment from the gel for cloning. There are several ways of doing this. The first is now the most widely used.

1. **Solubilizing or digesting the gel.** If the gel has been made using agarose with a low gelling temperature, then we can simply cut a slice out of the gel that contains the target DNA and melt the gel matrix by warming. Gels with a higher gelling temperature can be solubilized by treatment with chaotropic agents or digested with agarase. Once the gel has been disrupted, the DNA can be recovered by suitable solvent extraction or purification using silica.
2. **Diffusion.** A slice containing the DNA to be recovered is cut from the gel, crushed and soaked in buffer for several hours. Most of the DNA diffuses from the gel, which can be removed by filtration or centrifugation.
3. **Freeze–squeeze.** A slice containing the DNA is cut from the gel and frozen in liquid nitrogen. This breaks up the structure of the gel. It is then centrifuged through a glass wool plug. The plug stops the gel passing through, but allows the liquid from it (with dissolved DNA) to pass through.
4. **Electroelution.** There are several techniques involving electroelution. Here are three.
   (a) A slice is cut from the gel, sealed into dialysis tubing containing buffer and subjected to further electrophoresis. The DNA leaves the gel, but is retained in the buffer in the dialysis tubing.

(b) A slot is cut in the gel adjacent to and beyond the DNA of interest and filled with buffer, and electrophoresis continued. The DNA falls out of the gel into the slot containing the buffer, whence it can be removed with a pipette.

(c) A piece of a suitable membrane is inserted into the gel adjacent to the DNA and electrophoresis continued. The DNA sticks to the membrane. The membrane is removed and the DNA washed off with an appropriate solution. One suitable membrane is DEAE-cellulose, from which DNA can be removed by washing with a high-salt buffer.

1.7.4 Visualization

There are several ways of visualizing nucleic acids in gels. One of the easiest is to stain the gel with ethidium bromide and use fluorescence under UV light to locate the DNA. A related method is UV shadowing. Here, the gel is placed on a UV-fluorescent screen. UV light is shone onto the gel; this passes through the screen, causing it to fluoresce. DNA in the gel, however, will absorb the UV light, stopping it from reaching the screen. A non-fluorescing shadow will, therefore, be seen under regions of the gel that contain DNA. This technique is useful when dealing with single-stranded material, into which ethidium bromide does not intercalate very efficiently. It also has the advantage that it is not necessary subsequently to remove intercalated ethidium bromide, which may be difficult to do completely. Other dyes, such as methylene blue, can be used to locate DNA, although these are less sensitive. Radiolabelled DNA can, of course, be detected by autoradiography or fluorography.

1.7.5 Field-strength-gradient gels

A consequence of the way molecules are separated in a gel is that the greatest separation for a given difference in molecule length is obtained amongst the smallest DNA molecules, at the 'bottom' or anode of the gel. It is sometimes desirable to increase the separation in the upper part of the gel and reduce that in the lower. This can be done by having a non-uniform electric field strength (potential difference per unit length) in the gel, with a weaker field towards the bottom of the gel. The consequence is that as the molecules move further in the gel they experience a weaker field and so move more slowly. Molecules at the bottom of the gel, therefore, get separated less for a given difference in molecular length at the expense of those higher up.

Since the same current must flow through any cross-section of the gel, the gradient in field strength can be set up using a gradient in resistance, with lower resistance at the bottom of the gel. This is usually achieved in one of two ways. Wedge gels can be used, which are thicker and, therefore, of lower resistance at the bottom. Alternatively, buffer-gradient gels can be used, in which the concentration of buffer increases (and, therefore, resistance decreases) towards the bottom of the gel.

1.7.6 Pulsed-field gels

Pulsed-field gel electrophoresis (PFGE) can be used for separating much larger DNA molecules than those suitable for conventional electrophoresis, and can separate whole chromosomes. Very large DNA molecules (those above a certain threshold dictated by the pore size of the gel) can pass through the matrix only by threading through, oriented parallel to the direction of movement. Movement in any direction other than the one in which the molecule is 'pointing' is blocked by the gel matrix. Above that threshold size, all DNA molecules (if correctly oriented) will pass at the same speed. The principle of PFGE is to apply alternating electric fields across the gel in pulses, often at 90° to each other, although other angles can be used. Once the field position alters, the large molecules are unable to move until they become reoriented along the direction of the field. This takes longer for larger molecules. Therefore, provided the length of the pulse is greater than the time needed for reorientation, the time available for actual movement through the gel (and consequently the distance moved) will increase as the size of the molecule decreases. It is possible to resolve DNA molecules that are several megabases in size. There are several modifications to the technique, varying, for example, the position of the fields relative to one another, the relative strengths of the fields and the absolute and relative lengths of the pulses.

1.7.7 Blotting

It is often necessary to transfer DNA (or RNA) from a gel to a membrane as a solid-phase support. This is called blotting, and can be achieved using the arrangement shown in Figure 1.9. Buffer flows up from the reservoir via a wick, through the gel and the membrane over it, into the stack of paper towels above it. The nucleic acids are carried by the buffer, but trapped by the membrane over the gel. It is possible to assist the transfer electrophoretically or by use of a vacuum. The apparatus is then dismantled and the membrane removed. The nucleic acid can be denatured, if necessary (though it is usually denatured before transfer), and fixed to the membrane by heat treatment or UV crosslinking. If the membrane is then immersed in a solution containing labelled nucleic acids (a probe) complementary to some of the sequences on the membrane, the labelled probe will hybridize to the complementary material on the membrane. The position of the bound probe can then be visualized, for example, by autoradiography if the probe was radioactively labelled. The technique, therefore, allows one to work out which regions of a gel contained sequences corresponding to the probe. So, for example, it can be used to identify which fragment in the products of a restriction digest contains a sequence corresponding to a cloned gene used as a probe. If the gel contains mRNA, the technique will indicate the size class of mRNA (estimated by its position in the gel) that corresponds to the gene probe. A blot made

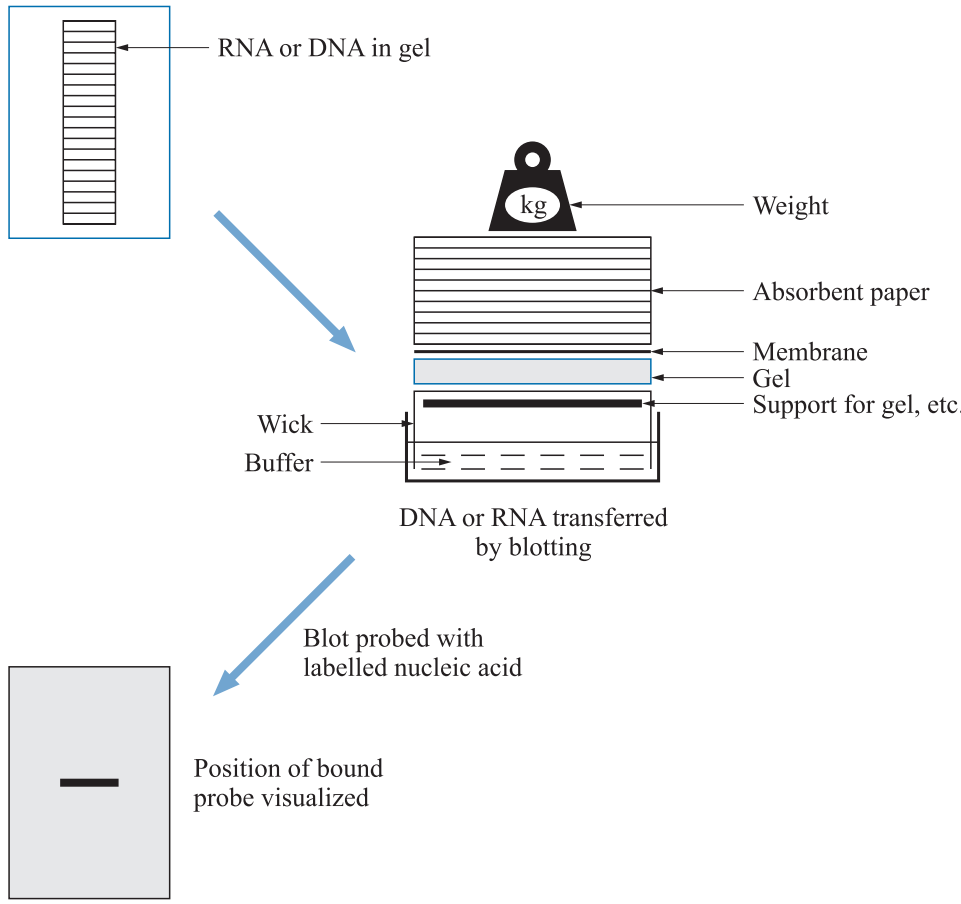


Fig 1.9 Blotting of a gel.

from a gel containing DNA is usually called a Southern blot, after its inventor Ed Southern. A northern blot is one taken from a gel containing RNA. It is possible to avoid transferring the DNA or RNA from the gel, and to probe the gel directly. Sometimes referred to as hybridization *in gello*, this approach is less sensitive than the conventional one. It can be useful, however, when the nucleic acid molecules in the gel are very large (tens of kilobase-pairs or more), which may reduce the efficiency of their transfer.

1.8 Oligonucleotide synthesis

Oligonucleotides of defined sequence are used widely, e.g. as primers for PCR or DNA sequencing, or for identifying complementary DNA molecules by hybridization. They are synthesized chemically, by the coupling of one nucleotide onto another, and although there are various approaches to synthesis, the principle is the same. Synthesis relies on the ability to block selectively the reactive groups on a nucleotide to limit the reaction to the addition of one nucleotide at a time. Thus, an individual round of coupling will result in the

addition of a single nucleotide. Unblocking one of the reactive groups then allows the addition of another nucleotide, and so on. Mixtures of nucleotides can be used in any round to allow the synthesis of a mixed population, in which different molecules can have different nucleotides at the corresponding position. Clearly, this technique is very repetitive, lending itself to automation, and a large variety of machines for it are available.

1.9 Microarrays

These are ordered collections of large numbers of different DNA sequences immobilized onto a solid-phase support, such as a microscope slide. Each position on the slide corresponds to a known DNA sequence. Very often, microarrays are designed to represent all the transcribed sequences of a target organism, and they can be used for a wide range of different purposes. For example, they can be used to study changes in transcript levels in response to a particular environmental or developmental change (Figure 1.10). RNA is prepared from the target organism before and after the change we are studying, and a DNA copy of the RNA population is made using reverse transcriptase. This newly synthesized DNA is also labelled using fluorescent dyes. The dyes Cy3 and Cy5, which fluoresce at different wavelengths, are widely used for this. Suppose we use Cy3 to label the DNA made using the RNA before the change we are studying and use Cy5 to label the DNA made from the RNA isolated after the change. The labelled DNA is then hybridized to the

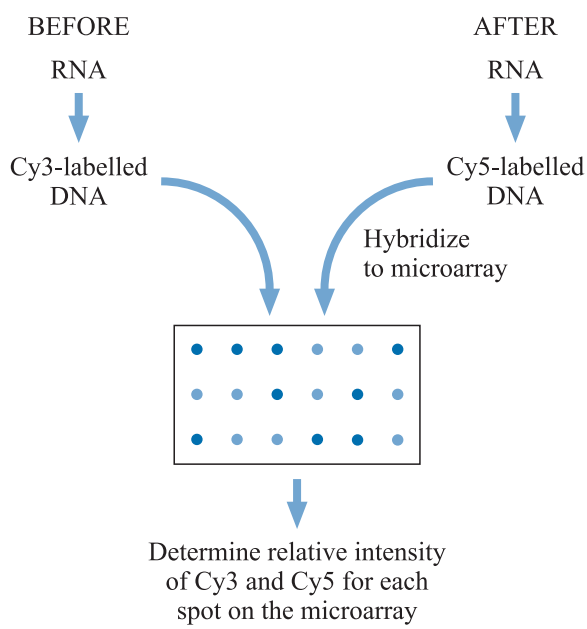


Fig 1.10 Microarray analysis.

microarray and the fluorescence due to the two different dyes is measured for each spot on the microarray. The more abundant a transcript of a particular gene is in the RNA sample, the stronger the signal we will see from the corresponding dye at the position in the microarray corresponding to that gene. So, in our example, a transcript that was more abundant in the second sample will give a stronger Cy5 signal, and a transcript that was less abundant in the second sample will give a stronger Cy3 signal. In this way, changes in transcript levels of a huge number of genes can be measured simultaneously. Microarrays can also be used in medical genetics to screen individuals suspected of having genetic deletions, by determining which DNA sequences in their genome are underrepresented by comparison with a sample of DNA from a control population.

Microarrays do not have to be made by immobilization of prepared, defined DNA sequences. They can also be made by synthesis of oligonucleotides *in situ* using photolithography technology from the microelectronics industry. These are sometimes called DNA chips.

Chapter 2

Polymerase chain reaction

2.1 The basic technique

2.1.1 The method

Later chapters will describe the techniques for the amplification of DNA sequences by propagation inside cells. However, it is often possible to amplify specific sequences more simply and quickly by a direct enzymatic process called the polymerase chain reaction or PCR. The basic procedure is outlined in Figure 2.1. In the simplest case, PCR amplification requires that we know a small amount of nucleotide sequence at each end of the region to be amplified. Oligonucleotides complementary to that sequence are synthesized, typically 20 or so nucleotides long. These oligonucleotides are used as primers for enzymatic amplification.

A reaction mixture is set up containing a sample of DNA that includes the region to be amplified, the primers in large molar excess, deoxynucleoside triphosphates (dNTPs) and a heat-stable DNA polymerase. The most common enzyme for this purpose is the Taq polymerase, which is a DNA polymerase isolated from the thermophilic bacterium *Thermus aquaticus*, which can be grown routinely in the laboratory at 75°C or more. This enzyme, which the bacterium uses for cellular DNA synthesis, has a temperature optimum of at least 80°C and is not readily denatured by the repeated heating and cooling cycles that we shall see are needed in the amplification process. There are many other thermophilic bacteria, and their polymerases can also be used, as discussed below.

The mixture is heated to a temperature sufficient to melt (i.e. separate the strands of) the sample DNA. After this it is cooled to a temperature low enough for the primers to anneal to the DNA. Of course, it is possible for the sample DNA simply to self-anneal, but this is rendered less likely by the large molar excess of the primers. The mix is then incubated at a temperature sufficient for the polymerase to synthesize a complementary strand to each piece of sample DNA, starting with the primers and using the dNTPs provided. We have, therefore, doubled the number of copies of the target region

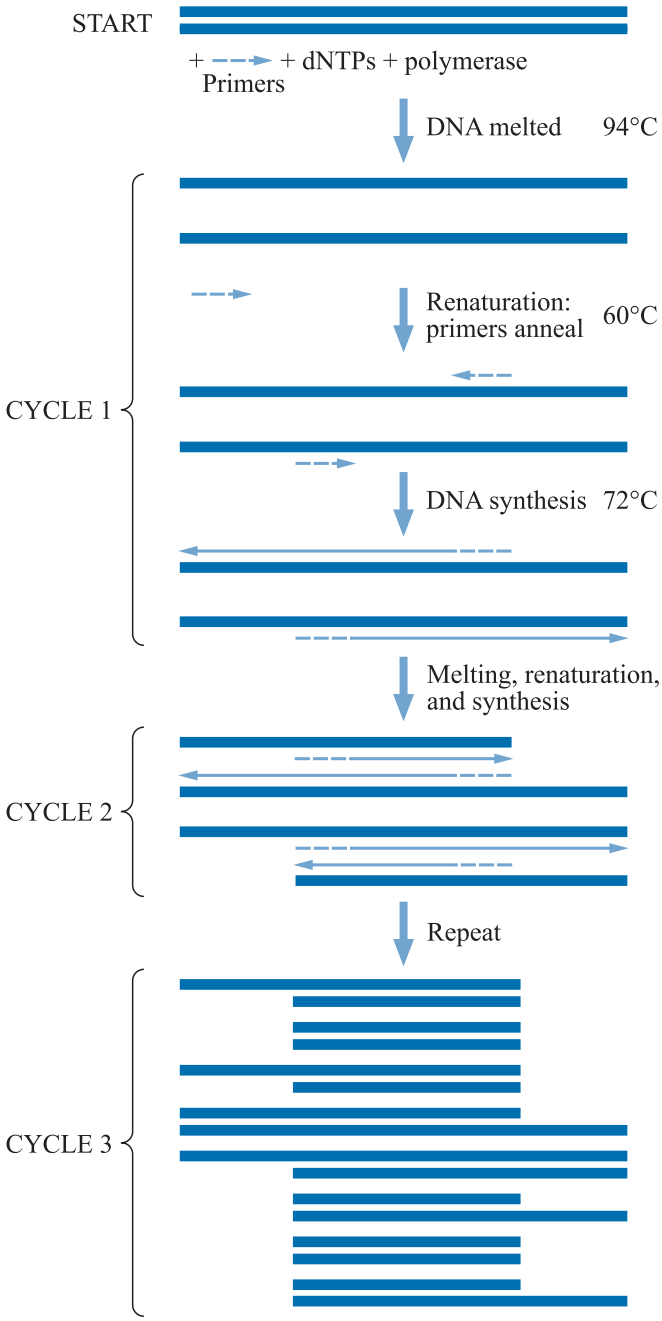


Fig 2.1 Polymerase chain reaction.

we are interested in. After this round of DNA synthesis, the mixture is heated for strand separation again, cooled again for primer annealing and incubated for another round of DNA synthesis. Each time the cycle is carried out there is a doubling of the number of copies of the target region between the primer annealing sites. A typical set of reactions might have an initial melting carried out for 5 min at 94°C, followed by 30 cycles each comprising melting

for 1 min at 94°C, renaturation for 1 min at 60°C and DNA synthesis for 1.5 min at 72°C, and then after the 30 cycles a final extended round of DNA synthesis for 10 min. This would result in a theoretical amplification of over 10^9-fold. Applied even to a single DNA molecule of 1 kbp, that would generate 1 ng of DNA, just about sufficient to see in a gel by staining with ethidium bromide.

Because of the repetitive nature of the process, it lends itself to automation, and many companies market thermal cyclers or PCR machines, which will carry out a series of programmed cycles of heating and cooling. Because the polymerase is thermostable and the primers were added in large excess initially, there is no need for further additions during the reaction; temperature control is all that is required.

Notice that not all the molecules generated will be of defined length. If this process is applied to genomic DNA (call this 'full-length' material), half the molecules will be full length after the first cycle, and the other half will start with a primer and have an undefined end, determined by how far the polymerase progressed during DNA synthesis (call these 'intermediate' molecules). In the next cycle, each full-length molecule will generate one full-length and one intermediate molecule. Each intermediate molecule will generate one intermediate molecule and one fully defined target molecule (begining and ending at the primer sites), as shown in Figure 2.1. As each cycle proceeds, the number of full-length molecules remains constant, the number of intermediate molecules increases arithmetically, and the number of target molecules increases essentially geometrically. After many cycles, therefore, gel electrophoresis of PCR products will indicate, to all intents and purposes, molecules all of a single size, corresponding to a single band in the gel. A full set of PCR cycles typically takes a few hours, assuming primers are already available, followed by gel analysis of the products. Synthesis of primers by commercial suppliers typically takes a day or so.

2.1.2 Polymerase types

As described above, the most commonly used enzyme is the Taq polymerase. It is generally produced by expression of the gene in *E. coli*. The thermostability of the Taq enzyme helps in its purification after expression in *E. coli*, since contaminating *E. coli* proteins can be inactivated by heating. The enzyme has 5′-3′ DNA polymerase and 5′-3′ exonuclease activities. It will polymerize about 50–60 nucleotides per second. However, the enzyme has a number of properties that may be disadvantageous.

1. **Taq polymerase has no proof-reading (3′-5′ exonuclease) activity.** Consequently about one nucleotide in 10^4 incorporated is incorrect and the individual products of PCR will be a heterogeneous population. If a mutation happens in an early cycle, then a large fraction of PCR products will have the altered sequence.

2. **Taq polymerase has relatively low processivity.** This means that it is likely to dissociate from the template before it has synthesized a long piece of DNA. Sometimes dissociation is caused by the incorporation of an incorrect nucleotide (i.e. one that is not base-paired to the template). The polymerase cannot correct the error, but it cannot readily elongate the strand being synthesized, as the 3′ end is not base-paired to the template. Consequently, the enzyme dissociates. The low processivity places a maximum size limit of molecules that can be amplified using this enzyme, which is typically 2-4 kbp.
3. **Taq polymerase is not fully heat stable.** It has a half-life of about 40 min at 95°C, which means there will be significant loss of activity over the 30 or so cycles used in a typical PCR experiment. It may, therefore, be necessary to add more enzyme during the course of an experiment.
4. **Taq polymerase incorporates an extra A residue.** This is incorporated on the 3′ end of the molecule synthesized, and is not template encoded. This overhanging A residue may help in cloning the products of PCR, but it depends on the strategy used.

A number of polymerases are available from other *Thermus* species. These include Tfl and Tth enzymes from *Thermus flavus* and *Thermus thermophilus* respectively. These generally do not have 3′-5′ proof-reading activity. Polymerases are also available from other genera of bacteria (including archaebacteria), and many of these enzymes have 3′-5′ proof-reading activity (which also means they do not usually add terminal nucleotides that are not template directed). Proof-reading enzymes include Tli and $Vent_R^{®}$ from *Thermococcus litoralis*, Pfu from *Pyrococcus furiosus*, and $DeepVent_R^{®}$ from *Pyrococcus* sp. GB-D. These marine bacteria generally grow at even higher temperatures than *Thermus aquaticus*, and the polymerases are more thermostable than the Taq enzyme. For example, *Thermococcus litoralis* can grow at temperatures up to 98°C, and the enzyme retains over 90% of its activity after incubation for 1 h at 95°C. *Pyrococcus* sp. GB-D was isolated from a depth of over 2000 m and grows at temperatures over 100°C. The $DeepVent_R^{®}$ enzyme derived from it has even greater thermostability than the $Vent_R^{®}$ enzyme.

2.1.3 Primer design

The strategy for primer design depends on the information available at the outset. In the simplest case, one may be attempting to amplify a piece of DNA whose sequence is known. It will then be possible to design primers that are completely complementary to the template region throughout their length. There are several computer programs available to suggest suitable primers, and some of the general guidelines are listed below.

1. **Length.** Short primers are more likely than long ones to match by chance to other sites in the template DNA than the intended annealing sites. The greater the complexity of the template DNA,

the more likely this is to happen. Thus, a short primer may offer sufficient specificity when amplifying using a simple template such as a small plasmid, but a long primer may be required when using eukaryotic genomic DNA as template. In practice, 20–30 nucleotides is generally satisfactory.

2. **Mismatches.** Primers do not need to match the template completely, although the 3′ end of the primer should be correctly base-paired to the template, otherwise the polymerase will not be able to extend it. It is often beneficial to have C or G as the 3′ terminal nucleotide. This makes the binding of the 3′ end of the primer to the template more stable than it would be with A or T at the 3′ end. Some mismatches in the body of the primer sequence may be allowed (although they will reduce the maximum temperature at which the primer remains annealed to the template). It is possible to have a run of several nucleotides at the 5′ end of the primer that do not anneal to the template. Thus, sequences can be incorporated at these ends that contain restriction endonuclease recognition sites (not present in the template) to facilitate subsequent cloning and manipulation of the PCR products.
3. **Melting temperature.** The temperatures at which the two primers can associate with the template should be relatively similar to ensure that they both bind at about the same time as temperatures are being lowered during annealing. The similarity of melting temperatures is likely to mean that the primers have a similar nucleotide composition.
4. **Internal secondary structure.** This should be avoided, or a primer may fold back on itself and not be available to bind to the template. As an intra-molecular reaction, self-annealing is likely to take place in preference to intermolecular annealing of the primer to the template.
5. **Primer–primer annealing.** It is also important to avoid the two primers being able to anneal to each other. Extension by DNA polymerase of two self-annealed primers leads to formation of a primer dimer as shown in Figure 2.2. These will be very efficient templates for amplification in subsequent rounds of PCR, as they are small.

2.1.4 Uncertainties in primer design

In many cases, the DNA sequence of the primer annealing site is not known with certainty. Here are some examples of when this may be the case.

1. **Primers are based on direct amino acid sequence.** We may have purified a protein of interest, determined the amino acid sequence of some or all of it and want to use that information to make primers to amplify the coding sequence. Because the genetic code is degenerate, we cannot predict the coding sequence with

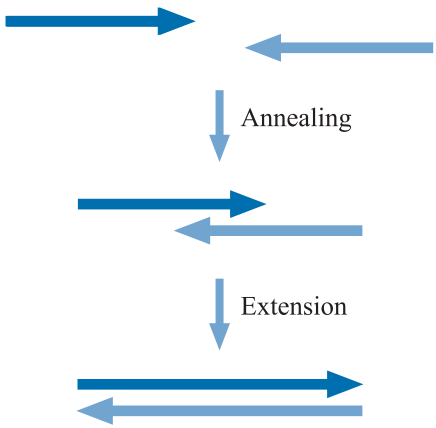


Fig 2.2 Formation of primer dimers. Annealing of two partially complementary primers allows extension and formation of a molecule that is a nearly complete dimer of the two primers.

certainty (except for amino acids such as methionine or tryptophan that, in the normal genetic code, have only one corresponding codon).

2. **Primers are based on related amino acid or nucleotide sequence.** It may be that we are interested in a family of genes (such as multiple copies of a gene in one organism, or homologous copies of the same gene in different organisms) and have sequence data from some members of the family and wish to use PCR to amplify additional sequences. Unless sequences are absolutely conserved between different members of the gene family (which is very unlikely), we will not know the sequence of the primer annealing sites with complete accuracy.

How can we solve the problem of not knowing the sequence of the primer annealing site with certainty? Many organisms use some codons for a given amino acid in preference to others, so we may be able to guess the codon likely to be used for a given amino acid if we have codon preference data available from other genes. If we are working from comparisons of amino acid sequences among a group of organisms and find that a particular amino acid is present in the great majority (but not all) of sequences at a particular position then we may choose to assume it will also be present in our target sequence. If we are still uncertain about what to include at a particular position in a primer, then we can include more than one nucleotide at that position. This is called a mixed site. Alternatively, we can use a nucleotide based on molecules, such as inosine, that have broader pairing capabilities than conventional nucleotides. Incorporating a mixed site at a position of uncertainty in a primer means that we can be confident the correct nucleotide sequence will be present. On the other hand, only a fraction of the primer molecules will have the correct sequence. Not only will the rest have a different sequence and be less likely to anneal at the correct site, they may be able to anneal at different sites, with the risk of the wrong region of the target DNA being amplified in PCR.

It is easiest to see how these primers might be designed by giving an example. Suppose we have the amino acid sequence of a protein from six different species and wish to use PCR to amplify some of the coding region from a seventh. Suppose we also know that the seventh species has a strong general preference for codons that have A or T in the third position, rather than codons for the same amino acid that have G or C in the third position. Comparison of the amino acid sequences identifies the following conserved region:

```
Species 1 MMFPGDAS
Species 2 MMFPGDAS
Species 3 MMFPGEAS
Species 4 MMLPGDAS
Species 5 MMFPGEAS
Species 6 MMFPGEAS
```

For amino acids 1 and 2, all species have methionine, for which there is only one codon, ATG. So we put ATGATG in the primer. The large majority of species have phenylalanine for the third amino acid, and one has leucine. There are six possible codons for leucine, so we may take a gamble and hope that the sequence we are interested in has phenylalanine, for which there are just two codons (TTT and TTC). Given the codon preference in the target species, we choose to put in TTT. At amino acid 4, all the sequences have proline, for which there are four codons, CCN (where N is any nucleotide). Given the codon preference, we can narrow this down to CCA and CCT, but no further. We may choose to put in a mixed site (A and T) at the third position of this codon, giving us CCA/T. The same argument applies to glycine at amino acid 5, so we use a mixed site and have GGA/T. At amino acid 6, the sequences are equally distributed between aspartate (GAT/C) and glutamate (GAA/G). The codons for these are similar; so, with this and the codon preference in mind, we can use GAT/A. For the seventh amino acid we have alanine with four codons (GCN). We could use GCA/T. However, we have serine as amino acid number 8, which has six codons. We may, therefore, choose not to have the serine represented in the primer and end with the alanine at number 7. If we also omit the third nucleotide from the alanine codon, we can end the primer with GC. This avoids the possibility of the 3′ nucleotide being incorrect, and also allows us to have C as the 3′ nucleotide, which is advantageous. The sequence of our primer for this region will therefore be

```
A T G A T G T T T C C A G G A G A A G C
                T     T     T
```

We have three mixed sites, each with two possibilities, so there will be a total of eight different sequences present.

2.1.5 Applications

Because PCR is automated and a typical set of cycles can be carried out on many separate samples simultaneously in a few hours, and starting amounts of DNA can be vanishingly small, PCR has a number of applications where the standard cloning techniques that we will cover later might be inappropriate, especially where speed and the number of samples to be processed are important or where the amount of DNA available is very limited. Here are some of the applications.

1. **DNA sequencing.** PCR in the presence of dideoxynucleoside triphosphates (ddNTPs), the chain-terminating inhibitors of DNA synthesis used for DNA sequencing, allows DNA sequencing reactions to be run successfully with very small amounts of template. This is known as cycle sequencing. It requires a specially developed enzyme that combines the properties needed for sequencing (especially processivity) with thermostability.
2. **Diagnostic.** PCR is useful as a diagnostic tool, e.g. in the identification of specific genetic traits or for the detection of pathogens or food contaminants. One of the first applications of PCR to genetic diagnosis was for sickle cell anaemia, allowing analysis to be completed within a day rather than the weeks taken by the conventional approach of hybridization analysis of DNA from cells. The sickle cell mutation in the beta globin gene destroys a restriction site, and the test involved PCR amplification of this region of the genome and analysis of the PCR products for the presence or absence of the restriction site. In general, PCR products can be analysed by
   (a) use of restriction enzymes, although it is rare for a mutation to create or destroy a restriction site;
   (b) determining whether an oligonucleotide probe specific for a particular allele is able to hybridize to PCR products;
   (c) electrophoresis, screening for mobility changes caused by sequence differences between a PCR product carrying a mutant allele and a wild-type molecule;
   (d) sequencing the PCR products directly.

   Tests for the presence of particular pathogens can be made using primers specific for the genomes of those organisms. This permits detection at extremely low levels. A similar approach can be used to detect DNA from contaminating sources in food.
3. **Forensic.** The ability to amplify DNA from regions of the genome that are highly polymorphic (and, therefore, which are variable between individuals) starting with samples containing very small amounts of DNA (e.g. single hairs or traces of body fluids, such as blood and semen) leads to applications in forensic work. A number of polymorphic regions have been used as targets for amplification, including: the D loop of mitochondrial DNA (mitochondrial DNA also has the advantage of being present at a high copy number per cell), which is variable between individuals

(much more so than the rest of the mitochondrial genome); the tandemly repeated minisatellites (also known as variable numbers of tandem repeats (VNTRs)) and microsatellites used in conventional genetic fingerprinting; and human leucocyte antigen (HLA) sequences. In one of the first widely publicized cases, the technique was applied in 1991 to the skeletal remains of a murder victim who had been wrapped in a carpet and buried from 1981 until the discovery of the bones in 1989. DNA isolated from these bones was heavily degraded, to fragments of less than 300 nucleotides, so analysis by conventional genetic fingerprinting techniques involving restriction digestion, Southern blotting and hybridization would not have been possible. The poor condition of the DNA also made PCR amplification of minisatellites impracticable, so amplification of shorter microsatellite regions composed simply of varying numbers of copies of the simple repeated sequence -CA- was used. PCR products from the victim could be matched with those from the putative mother and father, and the provisional identification of the body, which had been made earlier on the basis of facial reconstruction and dental records, was confirmed.

4. **Present-day population genetics.** The ability to amplify material rapidly from a large number of DNA preparations leads to applications in population genetics, allowing, for example, the determination of frequencies of particular alleles in a large collection of individuals. There are a number of PCR techniques that provide information from many parts of the genome simultaneously. One such is random amplification of polymorphic DNA (RAPD) analysis. This uses relatively short primers that will anneal to many different sites in the genome under study, producing many different bands when the PCR products are analysed electrophoretically. The similarities and differences between the band patterns generated from genomic DNA of different organisms provides information on their genetic relatedness. Another technique involves the amplification of a large set of restriction fragments of genomic DNA, whose sizes can be compared electrophoretically. Because of the similarity to conventional restriction fragment length polymorphism (RFLP) analysis, this technique is called amplified fragment length polymorphism (AFLP). PCR is also used in population genetic studies of microsatellites (or other tandemly repeated sequences), single nucleotide polymorphisms and transposons.

   A particular advantage of using PCR in population genetic studies is that, with appropriately designed specific primers, it may be possible to amplify DNA from one organism that cannot be separated from others, such as a particular bacterial strain in a mixed population. (Such primers will anneal to the target DNA from the organism of interest, but not to DNA from others.) This approach, therefore, can be used to study genetics of bacteria or other organisms that cannot be cultured axenically.

5. **Archaeology and evolution.** PCR can be used with old material as well as more recent samples, and it is often possible to amplify ancient (less-old material is sometimes called historic or stale) DNA from museum specimens and archaeological remains. Multiple copy sequences, such as mitochondrial DNA or chloroplast DNA, are a particularly useful target, although nuclear DNA is also retrievable in some cases. Comparison of polymorphic sequences from the ancient DNA with sequences observed today allows inferences to be made about the origins of particular populations or species. For example, the thylacine or marsupial wolf has been extinct for some time, although it was once a widespread species in Australia. Morphological data led to controversy over whether it was more closely related to a group of South American marsupial carnivores or to other Australian species such as the Tasmanian devil. Mitochondrial DNA was successfully amplified by PCR from skin and museum specimens collected in the late 19th and early 20th centuries. The sequence of the amplified DNA was determined, and it was interpreted as showing that the thylacine was more closely related to the Australian species than the South American species, and that the morphological similarities to the South American species resulted from convergent evolution.

   There has been much controversy over the length of time for which DNA will remain in amplifiable form in ancient material. There have been reports of amplification of DNA from material that is tens of millions of years old, but they have subsequently been shown to be due to contamination with microorganisms or other material (e.g. during handling of museum specimens or in laboratory extraction of DNA). When dealing with ancient DNA it is extremely important to follow appropriate authentication criteria, including suitable controls to detect any contamination. If the target material is degraded, then there may also be a possibility of generating chimeric sequences during PCR, by the jumping process outlined in Section 2.2.5.

2.2 Precautions and drawbacks

2.2.1 Size

The size of fragments that can be amplified is limited by the processivity of the polymerase used (see above). Using a mixture of polymerases that includes a proof-reading enzyme increases the size of product that can be obtained (up to 10 kbp or more), because incorrectly incorporated nucleotides can be removed rather than causing chain termination. The use of mixtures of enzymes to generate large products is sometimes called long-range PCR. In this case it may be necessary to adjust the time for DNA synthesis in the cycle times. If the template is heavily degraded, then it is less likely

to give large PCR products. Work with ancient DNA has shown an inverse correlation between the age of the template and the size of material that can be successfully amplified.

2.2.2 Amplifying the wrong sequence

PCR depends on the ability of the primers to anneal to the correct sequence, and this depends on the conditions of annealing (ionic concentration, temperature, etc.) and the actual sequence (or sequences if mixed sites are included) of the primers. It is possible for primers to anneal to the 'wrong' part of the target DNA, through chance complementarity. If this happens and the primers anneal in the correct orientation to each other (i.e. directing synthesis towards each other) and at sites that are not too far apart, then the result is the amplification of a sequence other than the desired one. It might be obvious that the wrong region has been amplified if the size of the amplification product (as estimated by gel electrophoresis) is different from that expected, but prior knowledge of the size is not always available. The possibility of incorrect annealing may be avoided by use of longer primers, which will be more specific in their annealing sites. Raising the temperature and adjusting the concentration of magnesium ions (which stabilize primer–template binding) can be used to increase the specificity of primer binding. Other ways of improving specificity are described in Section 2.3. As described above, it is also important to avoid using primers that can self-anneal or anneal directly to each other.

2.2.3 Contamination

Because of the extraordinary sensitivity of PCR, there is a particular danger of contaminating the DNA sample to be amplified with extraneous material. This is particularly important when using material containing only small amounts of DNA, as with archaeological work. Contamination might be of laboratory origin (e.g. from aerosols created by pipetting solutions containing related DNA sequences, including material amplified previously by PCR) or of external origin (perhaps by bacterial, fungal or human contamination of sample tissue).

Laboratory contamination can be minimized by precautions such as careful use and design of pipettes, and separation of the pre-PCR and post-PCR stages of an experiment into different rooms. Contamination from other sources can be reduced by careful handling and preparation of a sample before amplification (e.g. by removal of surface layers before DNA extraction, as these may be more heavily contaminated). Careful selection of primers may also help. For example, if obtaining mitochondrial DNA from mammoth bones, the use of species-specific primers would reduce the chance that contaminating fungal mitochondrial DNA will be amplified. Use of appropriate negative controls is also important, such as carrying out PCR using material from a 'dummy' DNA extraction with no sample tissue added, to indicate the presence of any PCR-amplifiable contamination in

buffers or in other reagents, or as aerosols. DNA from archaeological specimens is degraded, so that only short sequences of a few hundred base-pairs or less can be amplified. The ability to amplify long sequences may then be indicative of contamination with recent DNA, or the generation of chimeric gene products as discussed in Section 2.2.5.

2.2.4 Sequence heterogeneity

Amplification may give rise to a mixture of molecules of slightly different sequences. A mixture could arise for several reasons.

1. **Heterozygosity.** If the template DNA came from an individual heterozygous at the locus in question, each of the alleles present should be represented in similar quantities in the PCR products. This can easily be detected by direct sequencing of the PCR products, since it would give rise to two different signals at the same nucleotide position in the sequence output. However, if products are cloned before the sequencing (see Chapter 3), then the heterogeneity will not be detected if only a single cloned product is used for sequencing. (This is because individual clones are derived from single PCR product molecules.) Indeed, the heterogeneity may, by chance, go undetected even if several recombinants are used.
2. **Population heterogeneity.** If the template DNA came from several individuals rather than a single one, heterogeneity in the population may give rise to heterogeneity in the products. In this case, there may be many different alleles to identify.
3. **DNA damage and polymerase error.** Heterogeneity can also arise from damage to DNA before amplification, especially if the sample has not been carefully preserved. Therefore, this is particularly likely to be a problem with archaeological and forensic material. Damage includes oxidation or deamination of bases, and chemical cross-linking (within or between strands). As well as slowing down the polymerase (possibly resulting in the preferential amplification of undamaged, and perhaps contaminating, DNA), these changes may lead to the incorporation of incorrect bases during amplification. Thus, deamination of cytosine converts it into uracil, which would be read as thymine during DNA synthesis. If direct sequencing of amplified products is carried out, then this misincorporation may not be a problem, because the sample molecule that caused the misincorporation will be only a small proportion of the total number (unless the number of sample molecules is exceptionally small or the *same* site was affected in all the sample molecules). Consequently, the erroneous molecule will give a much weaker signal in sequencing than the genuine one does. However, sequencing of a cloned PCR product might lead to the generation of an erroneous sequence, as the cloned product derives from a single molecule and will give a single nucleotide sequence. If an error occurred during an early round

in the PCR process, then the majority of products would carry the erroneous sequence, and direct sequencing of PCR products might also generate the wrong sequence.

The polymerase may also make spontaneous errors not resulting from damage to the template DNA. The Taq polymerase, lacking a 3′-5′ proof-reading activity, gives an error rate of at least 1 in 10,000 nucleotides incorporated. Although this may seem quite low, a molecule of, say, 1 kbp would have a significant chance of containing a sequence alteration caused by polymerase error. Furthermore, an error that occurred in an early round of PCR will result in the presence of many incorrect molecules after several subsequent PCR cycles, as discussed above, with consequent problems for sequence determination.

Errors introduced by the polymerase, whether or not they result from template damage, can be expected to be evenly distributed over all three codon positions, whereas sequence heterogeneity (in coding regions) resulting from genuine allelic differences is likely to be concentrated in the third codon position and will have a tendency not to lead to amino acid substitutions.

2.2.5 Jumping PCR

When degraded DNA is amplified, it may be that any given sample molecule is not long enough to span the entire distance between the two priming sites. The result in the first round of synthesis would be extension of the primer to the end of a fragmented molecule, but not all the way to the second primer site. However, on a subsequent round of synthesis, the truncated amplification product may anneal to a different DNA fragment that contains the remaining region intact (see Figure 2.3). This would then allow synthesis of the full PCR product. This is called jumping PCR. So it is sometimes possible to generate PCR products that are longer than any individual template molecule. This can be advantageous when amplifying badly degraded DNA. On the other hand, jumping PCR has the distinct disadvantage of creating chimeric sequences that were not present initially. For example, if an individual is heterozygous at two sites within an amplified region, jumping PCR could generate molecules recombinant for these loci.

2.2.6 Interpretation

Interpreting the results of sequencing of PCR products from potentially damaged DNA (especially if it may have come from a mixture of individuals) necessitates answering the following questions:

1. Which nucleotides occur infrequently at a heterogeneous site? (The rare nucleotides may be the result of damage to the template.)

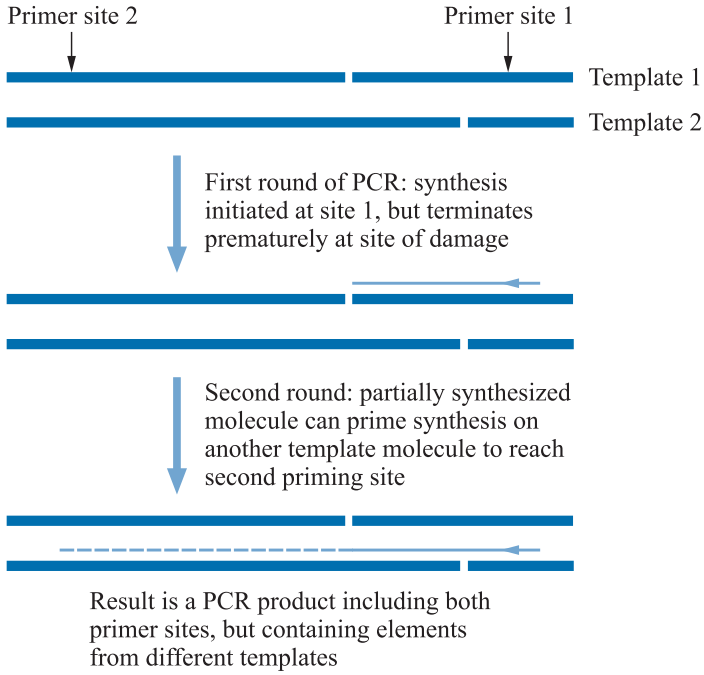


Fig 2.3 Jumping PCR. Although neither template molecule spans the entire distance between the priming sites, chimeric molecules spanning this distance can be generated by template switching in successive rounds.

PCR product				
	1	ACCGATTAGCTTAATATATGCATGGATTAGTTCCAGGTATTTA		
	2	G		
	3		G	C
	4		G	C
	5		G	C
	6			
	7			
	8			
	9		G	C
	10			
	11		G	C
	12		G	

Fig 2.4 Interpretation of PCR. The figure shows the result of sequencing 12 cloned PCR products. Only the differences between products 2–12 and product 1 are marked. Products 1, 6, 7, 8 and 10 are identical (Type A). Products 3, 4, 5, 9 and 11 are identical (Type B), differing from Type A at two sites and probably representing a different allele. Product 2 is unique and may represent damaged DNA, a polymerase error or a rare allele. Product 12 is also unique, but a hybrid between Types A and B. It could represent damaged DNA, polymerase error, another rare allele or jumping PCR between Type A and Type B molecules.

2. If, at a number of sites within the molecule, there are apparently genuine allelic differences (i.e. similar numbers of each nucleotide at a heterogeneous site):
   (a) Which combinations of sequences occur frequently (and are probably genuine alleles)?
   (b) Which combinations occur rarely (and might, therefore, be the result of jumping PCR)?

A theoretical example is given in Figure 2.4.

It is also important to be able to exclude the consequences of contamination, especially when dealing with ancient DNA.

2.3 Modifications

2.3.1 Improving specificity

One of the most frequently encountered problems is the annealing of the primers at the wrong location (or locations) to generate the wrong PCR product (or products). It may be possible to reduce this problem by raising the annealing temperature in the cycle conditions, or altering the ionic concentrations in the reaction mix, but there are a number of other approaches that we will now look at.

1. **Hot-start PCR.** As soon as the PCR reagents have all been mixed together, it is possible for the DNA polymerase to start synthesis. This may happen while the reaction mixture is being heated for the first time, and is at a temperature low enough to allow non-specific annealing of primer to template, generating a range of non-specific products. This problem would be prevented if DNA synthesis could not take place until the first cycle had reached its maximum temperature. This is the basis of hot-start PCR. In the simplest form, the DNA polymerase is not added to the reaction tubes until they have reached the DNA melting temperature of the first cycle. This is satisfactory where small numbers of samples are being processed, but not with large numbers. A more convenient approach is to have one of the reaction components in a form that is unavailable until an appropriate temperature has been reached. This can be achieved by incorporating the polymerase or the magnesium salt (needed for the polymerase to function) into wax beads. These beads melt at the appropriate temperature, allowing the reaction to start. A different approach is to have the polymerase inactivated at the start by complexing with an antibody. The antibody denatures at high temperature, allowing the polymerase to function.
2. **Touch-down PCR.** The annealing temperature used in conventional PCR is usually several degrees below the maximum at which primers can remain bound to template, to ensure stable binding. However, this use of a lower temperature permits a small amount of mismatching between primers and template, which may allow primers to bind to incorrect sites and generate spurious products. The effects of this can be reduced with touch-down PCR. In this, a high annealing temperature is used initially (at which even correct binding may not be possible). The annealing temperature is reduced in subsequent rounds. There will, therefore, come a point at which correctly matched primer–template annealing is just possible, but incorrect matching is not. DNA synthesis can therefore start. Although later cycles may be under less stringent conditions, the early cycles will have been carried out under the most stringent conditions and the desired products will be the most abundant.

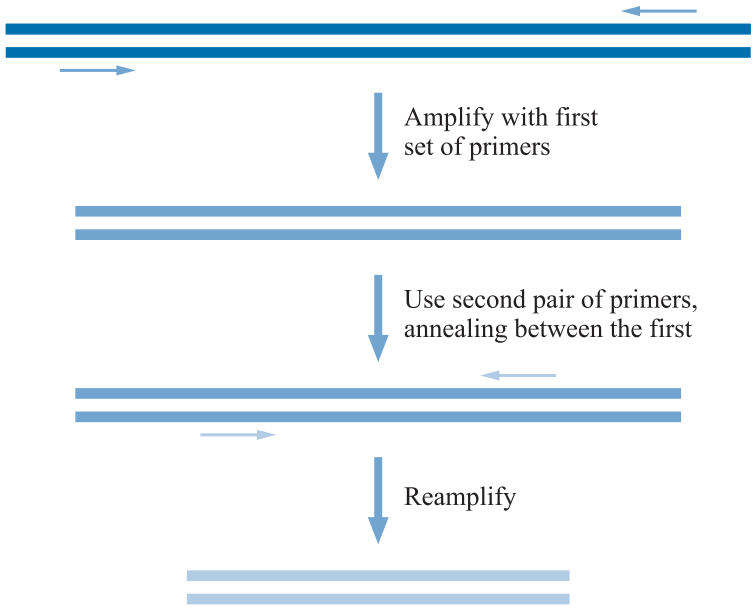


Fig 2.5 Nested PCR.

3. **Nested PCR.** Here, two successive PCRs are carried out. The first PCR uses a conventional template. The products of the first PCR are then used as the template for the second PCR, with primers that are designed to anneal within the desired product of the first PCR. Although the first PCR may generate some non-specific products in addition to the desired products, it is unlikely that the non-specific products will also contain annealing sites for both the primers used in the second PCR. Thus, only the desired products from the first PCR are likely to be suitable templates for the second. This is summarized in Figure 2.5.

2.3.2 Hard copies

Having amplified a specific piece of DNA, we may find it convenient to insert a copy of the products into a vector for subsequent cloning and maintenance by conventional means, as described later in this book. This is sometimes called 'making a hard copy'. The PCR product can be treated as a blunt-ended DNA molecule for cloning purposes. Another approach is to digest the PCR products with a restriction enzyme prior to cloning, as described in Chapter 3. If suitable restriction sites are not present within the DNA sequence to be amplified, then it is possible to incorporate them during amplification (see above). This can be done by inclusion of an appropriate sequence containing a restriction enzyme recognition site at the ends of the primers when they are synthesized. The primers will anneal to the target DNA, with the restriction site sequences unbound, as shown in Figure 2.6. The intermediate molecules that are formed will have the restriction site present at one end and all the target-length DNA molecules will have the sites

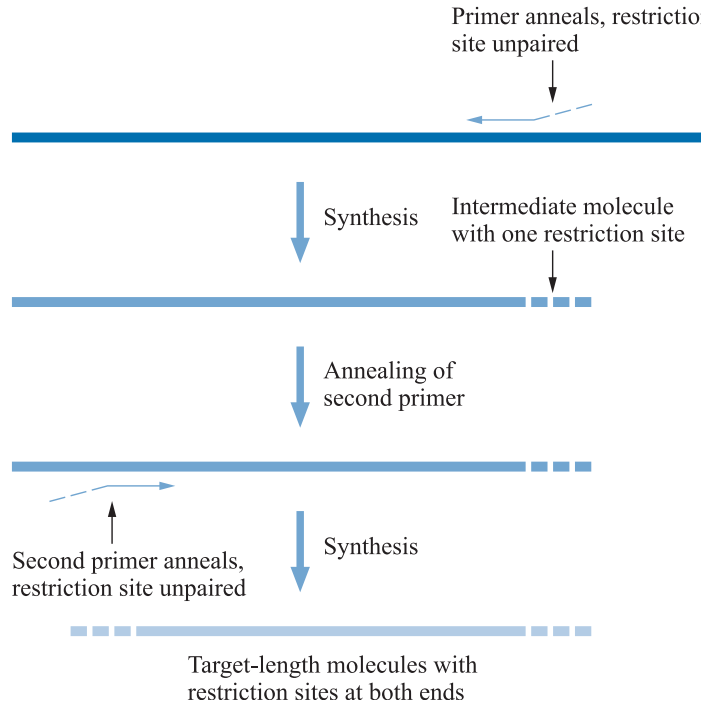


Fig 2.6 Addition of restriction sites in PCR. The dashed region indicates a part of the primer that contains a restriction site not present in the target DNA. For simplicity, only a single DNA is shown.

at both ends. The amplified molecules can then be cut at their ends with the appropriate enzyme(s) and the fragments cloned in the usual way. When designing primers for this purpose, it is usual to include a few nucleotides 5′ to the restriction site, because sites at the very ends of molecules may not be cut efficiently. The exact sequence of these nucleotides does not usually matter, as they will be lost during the cloning process. A third approach to making hard copies of PCR products relies on the fact that polymerases, such as the Taq enzyme, that lack proof-reading activity add an extra unpaired A residue at the 3′ end of the molecule they have synthesized. A PCR product made using such an enzyme will, therefore, have unpaired overhanging A residues at each end, allowing the molecule to be ligated to a linearized plasmid molecule that has unpaired T residues. Another approach is to incorporate topoisomerase I recognition sites into the PCR primers. Vector with covalently attached topoisomerase is then added, and the enzyme cuts the PCR product at the recognition site and joins it to the vector.

2.3.3 Inverse PCR

The result of the standard approach outlined so far is the amplification of sequence beween the primer annealing sites. However, it is also possible to arrange for the amplification of sequences *outside* the primers, in a technique called inverse PCR (IPCR). The approach is summarized in Figure 2.7. The sample DNA is first cut with an enzyme outside the region whose sequence is already known.

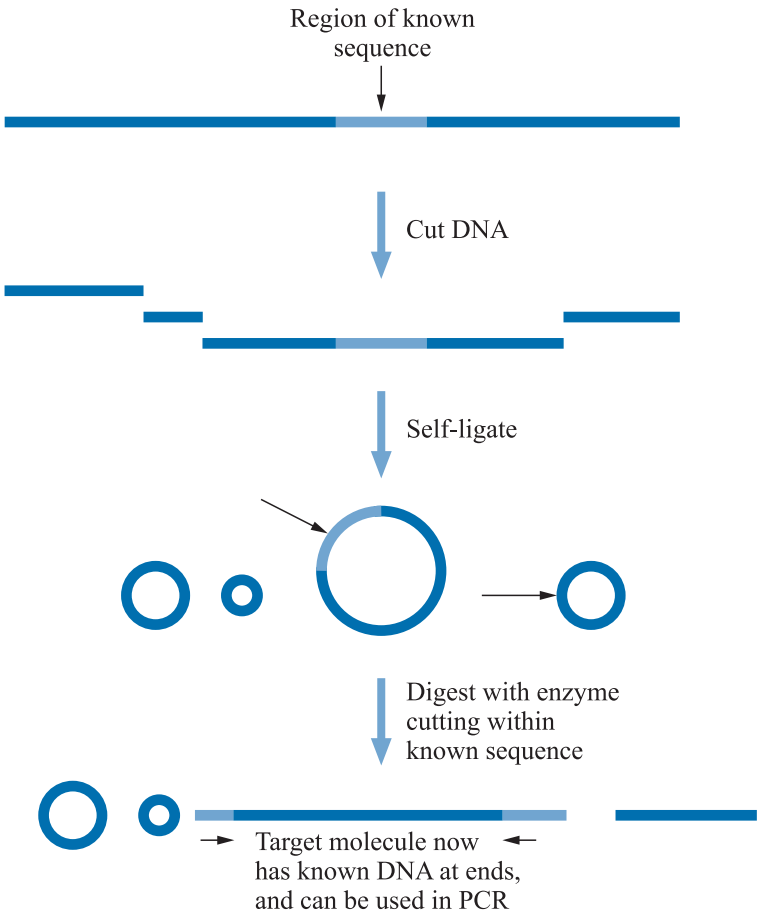


Fig 2.7 Inverse PCR (IPCR).

The resulting linear molecules are then circularized, by ligation under conditions that favour intramolecular reactions rather than intermolecular reactions (such as high dilution). A second restriction digestion is then done, using an enzyme cutting *within* the region of known sequence. The result is now that the first fragment containing this sequence has been turned 'inside out', leaving known sequence on the outside and the material that had previously been flanking it within. Primers complementary to the known sequence on the outside of the molecule can now be used to amplify the region of interest between them.

This technique is especially powerful when combined with transposon tagging (see Chapter 6). If a previously characterized transposon has inserted into a gene of interest, the gene can be amplified using IPCR and primers from the transposon sequence.

2.3.4 Reverse transcriptase PCR

It is often convenient to amplify RNA molecules, perhaps as a precursor to cloning them, or to estimate the abundance of a particular mRNA in a sample. This is usually done by having a round of reverse transcription, using a reverse transcriptase enzyme

and a single primer, to make a single strand of cDNA prior to the PCR itself. The primer for reverse transcription could be oligo-dT for general cDNA synthesis from polyadenylated messages, or it could be specific to a particular message.

2.3.5 In situ PCR

It is possible to carry out PCR using permeabilized tissue, such as thin sections on a microscope slide. This requires a specially adapted PCR machine to accommodate the slide. If the PCR product can be detected (perhaps by hybridization, also in situ), then this allows one to identify where in the tissue the target nucleic acid is located. Although this can be used to identify the location of particular DNA species in tissues, it is very often combined with reverse transcriptase PCR (RT-PCR) to identify the location of particular transcripts.

2.3.6 Quantitative PCR

It is possible to use PCR to estimate the abundance of a particular nucleic acid molecule in a sample. If we are interested in an RNA molecule it will be necessary to start with RT-PCR to make a cDNA copy. There are two main approaches to estimating the abundance of a molecule. One is to use a standard PCR and visualize the amount of the product of interest by gel electrophoresis, comparing it with suitable controls or standards. This is an end-point measurement of the amount of target sequence. The second approach is to quantify the reaction in real time (i.e. while the PCR is in progress). This can be done in two ways. In the first, a fluorescent, double-stranded DNA (dsDNA)-binding dye (such as SYBR green) is present in the PCR. As dsDNA product accumulates, the amount of fluorescence from the dye increases, and this can be detected. The experiment requires a PCR machine that is also equipped with a fluorescence measurement facility. This approach is adequate if the PCR generates the product of interest very specifically. Because the method simply detects dsDNA, it measures the amount of PCR product at a given time regardless of whether it is from the correct region.

The second approach to real-time PCR allows detection of a specific product, rather than dsDNA in general, and uses a specially synthesized probe oligonucleotide (Figure 2.8). This probe is designed to anneal within the region to be amplified and carries a fluorescent reporter dye at one end and a quencher at the other end of the molecule. If the quencher and the reporter are in close proximity (i.e. attached to the same oligonucleotide), then the quencher stops the reporter from fluorescing. During PCR, the probe will anneal to single-stranded DNA within the target region. When the polymerase meets the annealed probe, the 5′-3′ exonuclease activity of the enzyme degrades the probe, liberating the reporter from the quencher. Thus, the fluorescent reporter accumulates during the course of the PCR.

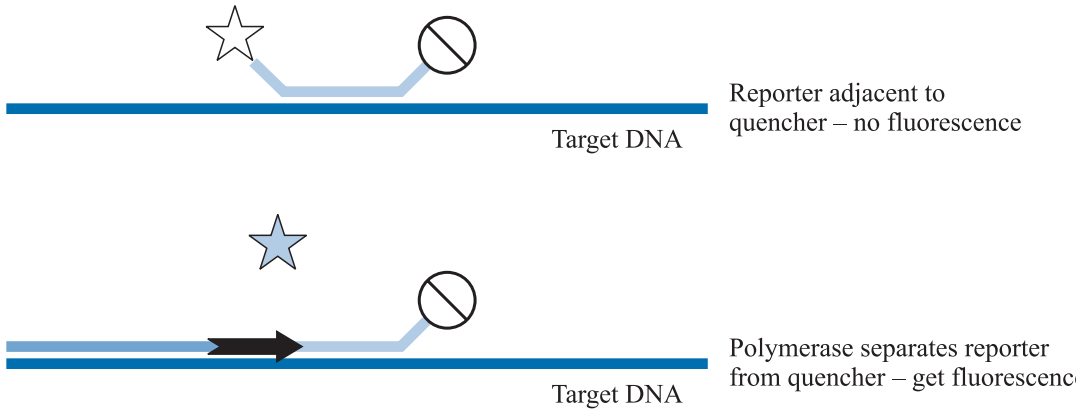


Fig 2.8 Real-time PCR. The figure shows the reaction monitored with a probe carrying reporter and quencher. Hydrolysis of the probe by the polymerase liberates the reporter, which fluoresces. ☆ = Reporter, ⃠ = Quencher.

An alternative method for specific product detection in real-time PCR uses two oligonucleotide probes that anneal to adjacent sites in the target DNA. One oligonucleotide is tagged with a molecule that absorbs light and the second with a molecule that is able to accept energy from the first and re-emit energy at a different wavelength. If the two oligonucleotides can both anneal to their target DNA, then the energy-absorbing and energy-emitting molecules are brought into close proximity and energy transfer is possible (Figure 2.9). If the oligonucleotides are not brought together in this way, then little energy transfer can take place. Thus, measurement of the amount of fluorescence from the second tag provides a measure of the amount of the target DNA.

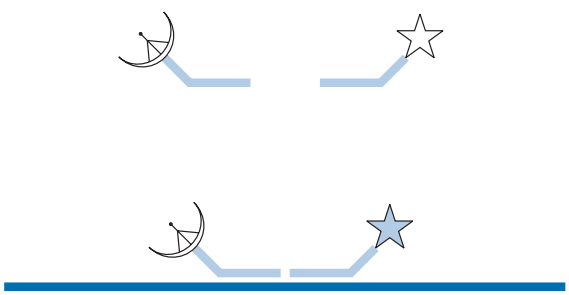


Fig 2.9 Real-time PCR. The figure shows the reaction monitored with separate probes that absorb light and fluoresce when brought into proximity. ◔ = Energy-absorbing molecule.

Real-time PCR is calibrated using samples with known amounts of template DNA. It is usually more accurate than end-point PCR. In the latter, the amount of product may be limited by the availability of reaction components, such as primers, rather than by the amount of target sequence. This is because the measurement takes place after all the cycles have completed. If limited by primer availability, then the amount of product is not directly related to the amount of target sequence. (It is, of course, also possible to take samples from a PCR during the course of the reactions and measure the amount

of product at intermediate stages by electrophoresis. This avoids this problem, but it is very time consuming.)

There is a danger of confusion in the terminology here. The term RT-PCR is used as an abbreviation for reverse transcriptase PCR. However, RT-PCR is also used by many people as an abbreviation for real-time PCR. The potential for confusion is exacerbated by the fact that measurement of RNA levels by real-time PCR requires an initial reverse transcription step! A simple solution is to use RT-PCR only to mean reverse transcriptase PCR, and to refer to the quantitative, real-time, methods as qPCR (or quantitative PCR).

2.3.7 Mutagenesis

Perfect matching between the primer and the target DNA is not necessary, provided allowance is made for this in determining the conditions for primer annealing. PCR can, therefore, be used to engineer specific mutations in the amplified sequence by constructing a primer that corresponds to the mutated sequence rather than the original. The use of this is described in more detail in Section 7.3.

2.3.8 Asymmetric PCR

By reducing the amount of one of the two primers, it is possible to arrange for preferential amplification of one of the strands, resulting in a preparation of single-stranded DNA, which has a number of uses in molecular biology. (These include DNA sequencing, in which case the other amplification primer could serve as a primer for the sequencing reactions.) Preferential amplification of one strand in this way is known as asymmetric PCR. Single-stranded material can also be obtained by tagging one of the primers (e.g. with biotin) and then using an affinity purification method (streptavidin-based, if biotin was used) to separate the tagged strand.

2.3.9 Anchored PCR

Anchored PCR is applied when only one piece of sequence (and therefore one priming site) for the region of interest is known. The aim is to attach the region to be amplified to a piece of known sequence and then to use that as the second priming site. There are two ways in which this can most easily be done. One is to fragment the sample DNA and ligate it to molecules of known sequence, such as a vector (Figure 2.10). This known sequence is used as the basis for designing one of the two PCR primers. The second method is to add tails enzymatically to the sample DNA or the molecules produced after the first round of synthesis. For example, a run of, say, G residues can be added by treatment with terminal deoxynucleotidyl transferase and 2′-deoxyguanosine 5′-triphosphate (dGTP). (This is an unusual example of DNA synthesis that does not require a template.) Oligo-dC can then be used as a primer, because it will anneal to

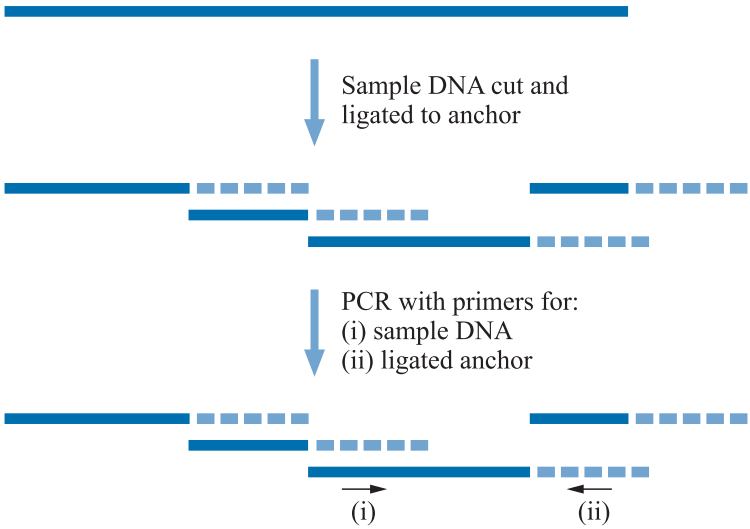


Fig 2.10 Anchored PCR. The dashed line represents a double-stranded molecule of known sequence (an anchor) that provides one of the two priming sites for PCR. The primers are shown by (i) and (ii).

the oligo-dG. Combined with RT-PCR, this forms the basis of Rapid amplification of cDNA ends, (RACE) described in Section 5.3.

2.3.10 Emulsion PCR

In a conventional PCR, the reactions are carried out inside plastic tubes. It is possible to incorporate all the reagents inside lipid droplets and carry out PCR on a much smaller scale. This has certain advantages. It is possible to increase and decrease the temperature of small droplets very quickly. In addition, if each droplet contains a single template molecule at the start, then all the products in an individual droplet result from the amplification of a single template molecule. The method is also called droplet PCR.

2.3.11 Isothermal amplification

The repeated heating and cooling required by PCR limits how quickly the process can be carried out. In addition, the cost of a PCR machine is not trivial. Loop-mediated isothermal amplification (LAMP) has been developed, which allows templates to be amplified at a constant temperature (typically around 65°C). It uses a DNA polymerase with strand-displacing activity and avoids the need for heating to high temperatures. There is particular interest in the use of this method for the detection of pathogens outside of specialist laboratories, as it is rapid and avoids the need for expensive PCR machines.

Chapter 3

Simple cloning

3.1 The basic experiment

3.1.1 Cloning using a simple plasmid

Now that we have surveyed the enzymes and techniques that are used in cloning, we can take an easy example and see how they are applied. One of the simplest exercises is inserting DNA into *E. coli* in a plasmid such as pUC18 (Figure 3.1). Bacterial plasmids are circular dsDNA molecules capable of replicating in bacterial cells. Insertion of DNA into the plasmid allows the DNA to be propagated in a host cell. Molecules that are used for propagating DNA in this way are called vectors. The pUC18 plasmid is one of a series of similar vectors and carries two genes of particular importance. One confers resistance to the antibiotic ampicillin (Figure 3.2), which blocks the cross-linking of the polysaccharide chains in the bacterial cell walls. The lack of cross-linking weakens the bacterial cell wall and the cells eventually lyse. The resistance gene (*bla* or amp^R) encodes a beta-lactamase enzyme that hydrolyses the beta-lactam ring of the antibiotic, inactivating it. The other gene (*lacZ′*) encodes part of

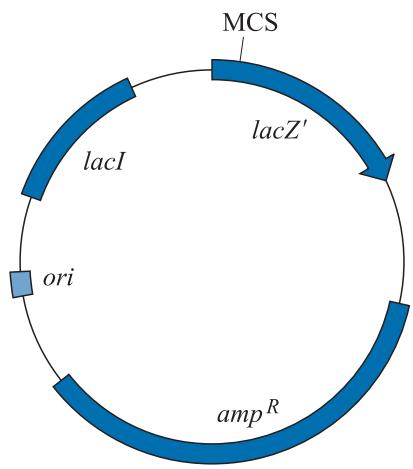


Fig 3.1 Generalized pUC plasmid. It contains an ampicillin resistance gene (amp^R), part of a beta-galactosidase gene (*lacZ′*), the Lac repressor gene (*lacI*) and an orgin of replication (*ori*). The multiple cloning site (MCS) is located in the beta-galactosidase gene.

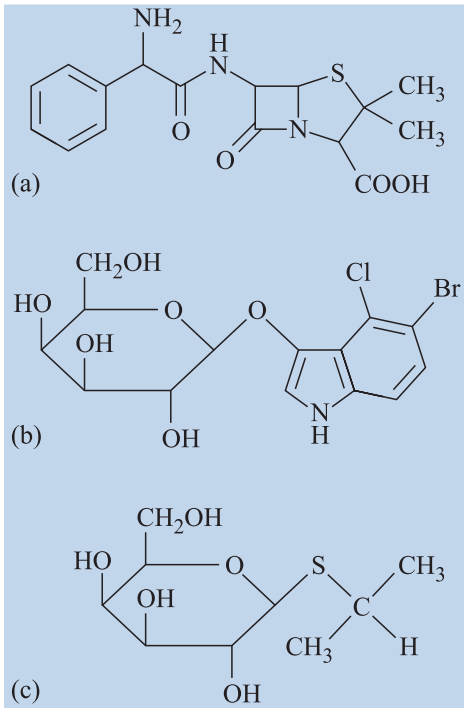


Fig 3.2 Structure of ampicillin (a), X-gal or 5-bromo-4-chloro-3-indolyl-β-D-galactoside (b), and IPTG or isopropyl-β-D-thiogalactoside (c).

a beta-galactosidase enzyme, which normally cleaves disaccharides, such as lactose, into monosaccharides. It can also cleave an artificial substrate 5-bromo-4-chloro-3-indolyl-β-D-galactoside (also known as X-gal), shown in Figure 3.2, to liberate a blue dye. For this reason, the substrate is said to be chromogenic. We will see later why the gene for the whole beta-galactosidase enzyme is not used, but for the moment we can treat the plasmid as though it contained the entire gene. As well as the amp^R and *lacZ′* genes, the plasmid has some less immediately obvious features that we will look at in more detail later, such as an origin of replication. It also has a number of recognition sites for restriction endonucleases, and many of these are located in a region of the *lacZ′* gene known as the multiple cloning site or polylinker.

Suppose we wanted to clone genomic DNA from the organism we were studying into the *Bam*HI site of pUC18, located in the multiple cloning site in the *lacZ′* gene. (We do not have to use the *Bam*HI site. We could use any of the other restriction sites in the multiple cloning site.) The procedure would be as follows. The first stage is to purify genomic DNA from the organism of interest and cut it with the enzyme *Bam*HI. We would also purify some pUC18 and cut it with the same enzyme. Then the *Bam*HI-digested plasmid and the *Bam*HI-digested genomic DNA are mixed and DNA ligase in a suitable buffer is added. *Bam*HI ends will anneal to *Bam*HI ends and the ligase will seal the nicks. That will generate a large number of types of molecule, which may be linear or (topologically) circular.

Linear molecules will not be stably propagated in later stages of the experiment, so we will not consider them further. Different circular molecules that may be formed are shown in Figure 3.3. Intramolecular ligation of plasmid DNA will regenerate the original circular pUC18 plasmid. Intramolecular ligation of genomic DNA will produce circular molecules of genomic DNA, with no plasmid DNA sequences present. A wide range of intermolecular ligations can take place. Two plasmid DNA molecules can be ligated, in a head-to-tail or head-to-head configuration. Two genomic DNA molecules could be ligated, again in either orientation. Most important, a genomic DNA molecule can be ligated to a plasmid DNA molecule. (Of course, a huge range of ligations involving three or more molecules can also take place, but they will be rarer than unimolecular or bimolecular ligations. How rare depends on the concentrations of the species involved.)

The molecules that contain new combinations of sequences, not present before, are termed recombinants. Note that the insertion of DNA into the *lacZ′* gene will destroy the function of that gene, as it disrupts the coding region for the LacZ′ protein; therefore, it is an easy matter to work out what properties each of the plasmid types shown in Figure 3.3 will confer. It is possible that insertion of a short piece of DNA into the *lacZ′* gene might only partially inactivate the gene function (or not affect it at all) if the reading frame were not terminated or shifted. This is relatively uncommon, but the possibility should always be kept in mind.

The next stage is to put the ligation products back into *E. coli*. For an experiment such as this, it would usually be done by chemically induced transformation, or electroporation (Sections 1.5.2 and 1.5.3). An important point is that the uptake of DNA by *E. coli* is rather inefficient, in that the probability of any individual bacterial cell taking up a piece of DNA is very low. For simple transformation procedures, of 10^9 cells treated with 1 μg of supercoiled plasmid DNA, only 10^5 or so will successfully take up a molecule. We say there is a transformation frequency of 10^5 colonies/microgram of DNA. This is low, given that 1 μg of pUC18 contains of the order of 10^{11} molecules. The cells are now plated on agar containing, in this example, ampicillin, together with an inducer of the *lacZ′* gene (isopropyl-thiogalactoside (IPTG), Figure 3.2) and the chromogenic substrate X-gal. The cells that have not taken up a plasmid will be sensitive to the ampicillin and die after a few generations. Those that took up genomic DNA only will also die. Those that took up pUC18 DNA will have acquired an ampicillin resistance gene and be able to grow on the selective medium. They will, therefore, form isolated colonies on the agar if plated at a suitable dilution. Thus, all the members of a colony are derived from one cell and should have acquired a single plasmid. (Because the probability of a cell taking up one DNA molecule is low, the probability of a cell taking up two might be expected to be very low indeed. Actually, that is likely not to be the case. Although only a tiny fraction of cells is able to take up DNA,

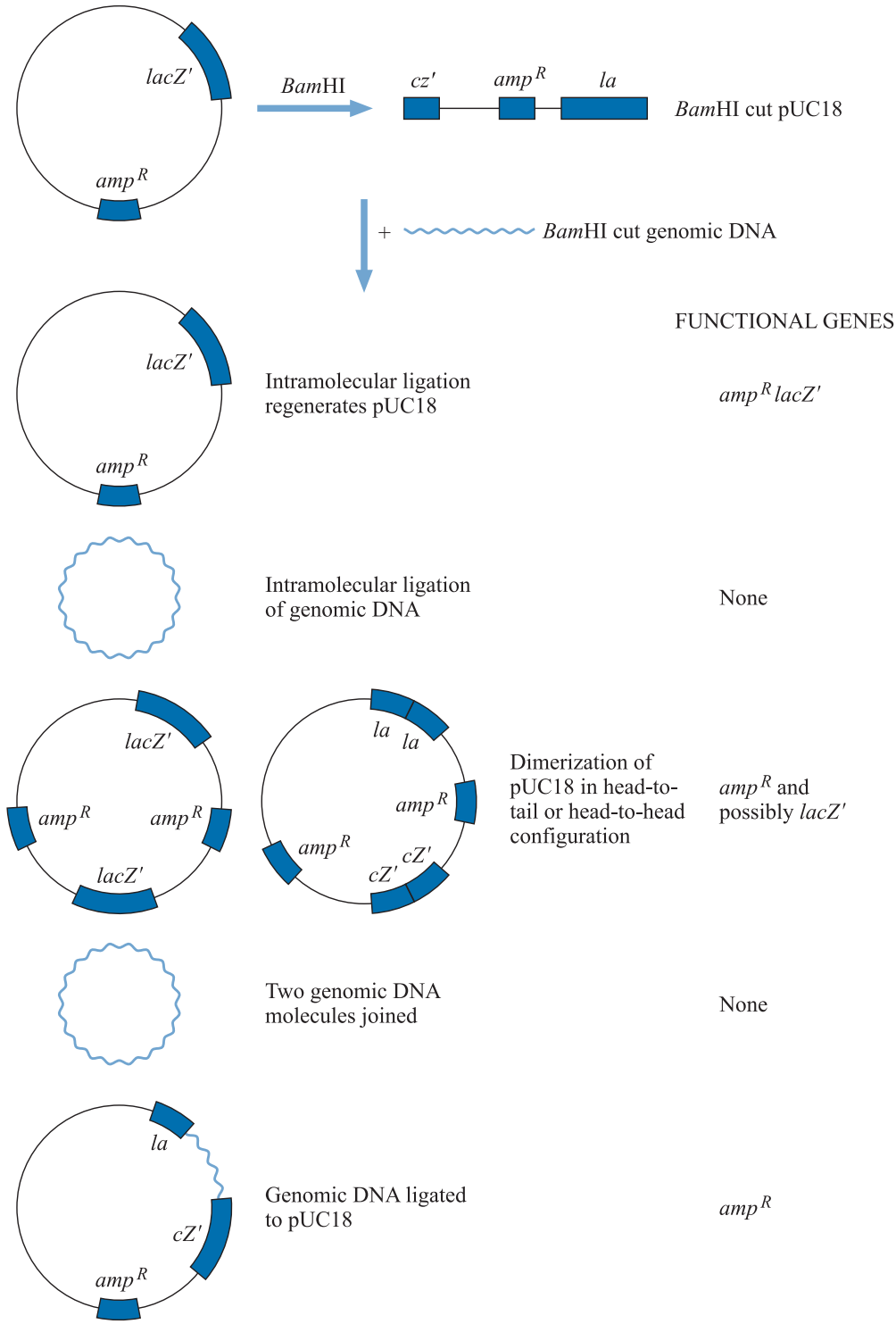


Fig 3.3 Ligation of *Bam*HI-cut genomic DNA into a *Bam*HI-cut vector. The figure shows the regeneration of vector, the circularization of genomic DNA and the ways in which two separate molecules (vector or genomic DNA) can be joined. The functional genes present on each molecule after ligation are indicated.

those cells that do are quite efficient at DNA uptake and may take up more than one piece. So a low, but significant fraction of colonies may contain two different plasmids. Over time, these may segregate, with some cells containing one plasmid and the rest containing the other.)

The problem now is to distinguish the colonies that have acquired genomic DNA from all the others. That can be done in this example by using the *lacZ′* gene. Colonies that have acquired pUC18 molecules without an insert of genomic DNA will have an intact *lacZ′* gene (which will be induced by the IPTG). They will, therefore, be able to break down the X-gal and be coloured blue. The colonies that have acquired molecules with a genomic DNA insert will have a *lacZ′* gene that has been inactivated by the insertion. Thus, these colonies will be unable to metabolize the X-gal and will be the usual white (actually off-white) colour. The head-to-tail dimeric pUC18 molecules will also give blue colonies, and the head-to-head dimers will not replicate stably in the host cell. The relative abundances of blue and white colonies will depend on factors such as the relative proportions of genomic and plasmid DNA in the initial ligation. We are interested in the white colonies. We can confirm that plasmids in the white colonies contain an insert by growing up the bacteria, purifying plasmid DNA and cutting it with a restriction enzyme. *Bam*HI would be the easiest in this case, and should generate two sizes of molecule: plasmid and inserted genomic DNA. Plasmids from blue colonies, where there is no insert, will give a single size of restriction fragment, corresponding to the plasmid DNA. The same would be seen for the head-to-tail dimeric plasmids. (An alternative way to differentiate whether plasmids have an insert, and to measure the size of the insert, is to do PCR on DNA from the colonies using primers that anneal on each side of the cloning site. This can be done directly with cells, without needing to purify the DNA, and is called colony PCR.)

The outcome of this experiment is a large number of *E. coli* colonies, each carrying a plasmid containing a *Bam*HI fragment of genomic DNA. This is useful in a number of ways. For example, to get a lot more of each piece of genomic DNA, it is just a matter of growing up a larger quantity of *E. coli*, extracting the plasmid and, if necessary, separating the insert from the plasmid. That is much easier than extracting the DNA from more of the organism(s) we used in the first place. Also, each plasmid insert is a single, defined piece of DNA. To separate a single, defined piece out of the total genomic DNA each time one wanted it would be too laborious. This propagation of individual, defined pieces of DNA in a suitable organism is cloning. The generation of recombinants in this way by randomly cloning fragments of the total DNA from an organism is called shotgun cloning, and the random collection of *E. coli* colonies containing the recombinant DNA molecules is called a library. Because total genomic DNA was used, it is called a genomic library. An ideal genomic library would contain all the sequences present

in the original organism's genome, and would then be described as representative. A library constructed in the way we have just described would probably not be very representative. For example, small DNA fragments tend to be cloned more efficiently than large ones, so our library would be underrepresentative of the large fragments. Another problem with the library in our example is that any gene that contained a *Bam*HI site could never be rescued intact, because *Bam*HI digestion was used to generate the fragments to be cloned. It would be better to use some other way of generating the fragments (see Section 5.2).

3.1.2 Use of the *lacZ′* indicator system (alpha-complementation)

In this example, we said that the pUC plasmid contained only a part of the *lacZ* beta-galactosidase gene. Beta-galactosidase is rather a large protein (116 kDa), and it is more convenient to have just a portion of the gene on the plasmid to keep the plasmid as small as possible. This portion encodes the first 146 amino acids and is sometimes called the *lacZ′* minigene. In addition to the *lacZ′* gene and the ampicillin resistance gene, the plasmid also contains a *lacI* gene for the Lac repressor protein (Figure 3.1). This keeps the minigene repressed, except in the presence of an inducer such as IPTG. This may be particularly important if the sequences inserted into the minigene are toxic if expressed at high levels. The rest of the *lacZ* gene is contained within the host, often on an F′ plasmid. In fact it is a version of the gene called the M15 deletion, which lacks codons 11–41. The polypeptide produced from this fails to tetramerize, and this tetramerization is needed for enzymatic activity. However, in the presence of the minigene product (amino acids 1–146), assembly can take place to produce a molecule with low but detectable beta-galactosidase activity. This intragenic complementation is called alpha-complementation. The principle of alpha-complementation is summarized in Table 3.1. In essence, it is just a way of reducing the size of the vector needed. The pUC vectors were developed at the University of California (hence their name). We will now go back to the basic procedure we have described and look in a little more detail at the vectors and hosts.

Table 3.1 Alpha-complementation

	LacZ residues 1–146 present?	M15-deleted LacZ present?	Result (+IPTG+X-gal+Amp)?
Host	No	Yes	Amp^S, no LacZ; no colonies
Host + pUC plasmid	Yes	Yes	Amp^R, LacZ; blue colonies
Host + pUC plasmid (with insert)	No	Yes	Amp^R, no LacZ; white colonies

3.2 Vectors, transformation and hosts

3.2.1 Vectors

As outlined above, vector is the name given to DNA molecules into which foreign DNA is inserted for subsequent propagation in a host cell. A vector should have several features:

1. **An origin of replication**. Without an origin of replication, the vector could not replicate, and when the cells divide after taking up the vector molecule, only one of the daughter cells would retain it. Therefore, we would never get a colony of transformed cells. The pUC18 origin of replication came ultimately from a plasmid in a clinical bacterial isolate denoted as pMB1.
2. **A selectable marker**. This is needed to distinguish cells that have taken up the vector from those that have not. In the example we have just looked at, the selectable marker was ampicillin resistance. The ampicillin resistance gene derives from a transposon (a piece of DNA able to move, or transpose, to different places in the genome) from another plasmid, pRSF2124.
3. **Suitable single restriction sites.** In the example we considered, pUC18 had just one recognition site for the enzyme *Bam*HI, which was used in the cloning. Had there been more than one *Bam*HI site in pUC18, the vector would not have been suitable, because cutting with *Bam*HI would have cut it into more than one piece. Reassembly of the vector together with an insert during the ligation would have required a trimolecular reaction at least, which would have been very unlikely. Note, though, that the same enzyme does not necessarily have to be used for cutting the vector and insert (see Section 1.2.1). If the insert molecules had been generated by *Sau*3A digestion, we could still have used *Bam*HI to cut the vector, as *Bam*HI ends are compatible with *Sau*3A ends. Cutting the vector with *Sau3*A would not have been feasible, as there are too many sites for *Sau3*A in the vector. Note also that restriction sites located in indispensable genes in the vector are unsuitable for cloning, because insertion of DNA there would be likely to destroy gene function and, therefore, vector viability. We will see that this is particularly important when we consider vectors based on bacterial viruses in Chapter 4.
4. **Suitable size.** To some extent, having a suitable size is a corollary of having suitable single restriction sites. A restriction enzyme cleavage site that comprises six nucleotides will occur on average approximately once every 4^6 bp (i.e. every 4 kbp or so). So a vector that was much larger than 4 kbp might be expected to have several sites for a given enzyme, and cutting the vector would reduce it into several pieces that would be unlikely to be correctly assembled in a ligation reaction. It is possible to remove restriction sites, and methods for doing this are described in Chapter 7. Simply removing excess restriction sites may not eliminate

the problem, though. Large DNA molecules are very susceptible to physical shearing, even in the simple act of pipetting, so they are always difficult to handle. Note that the use of the alpha-complementation system allowed the size of the pUC plasmids to be reduced from what would otherwise be necessary to encode beta-galactosidase.

5. **Markers for DNA insertion.** In the example we studied, the insertion of DNA into the vector can be detected by inactivation of the *lacZ′* gene, which in turn could easily be assayed by plating cells onto medium containing an inducer and a chromogenic substrate as well as ampicillin. In some older vectors, an additional round of plating was needed to distinguish recombinants from non-recombinants. An example of this is the vector pBR322, which contains an ampicillin resistance gene and a tetracycline resistance gene (Figure 3.4). There are cloning sites in both genes. Suppose we had used the *Bam*HI site in the tetracycline resistance gene for insertion of DNA. After transformation, we would plate the *E. coli* cells on medium containing ampicillin to select for the acquisition of a plasmid. We would then plate small samples from ampicillin-resistant colonies onto medium containing tetracycline. If DNA had been inserted into the *Bam*HI site, then the tetracycline resistance gene would be inactivated, and cells containing recombinant plasmids would be sensitive to tetracycline. Cells that had acquired a plasmid without inserted DNA would be resistant to tetracycline as well as ampicillin. This need for two rounds of plating was one of the reasons for the pBR322 plasmid being superseded by the pUC and other related plasmids as a routine cloning vector. In some of the first plasmids used for cloning, the cloning site was not in a functional sequence at all. The only way to detect insertion, therefore, was to isolate plasmid DNAs, digest with a suitable enzyme and examine the

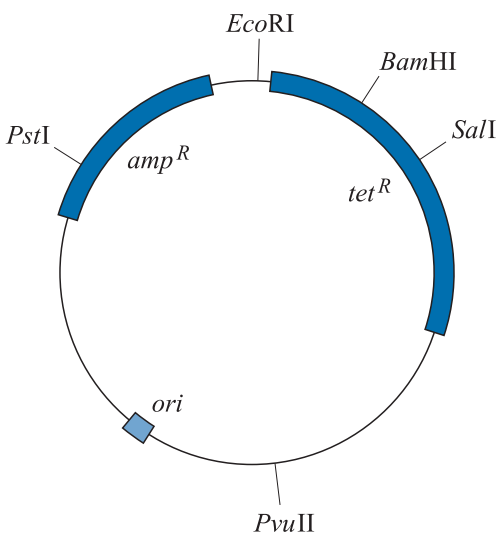


Fig 3.4 The plasmid pBR322 (4.3 kb). The figure shows genes for ampicillin resistance (amp^R), tetracycline resistance, (tet^R), the origin of replication (*ori*), and a selection of unique restriction sites.

fragments produced electrophoretically. That was potentially very tedious, especially if intramolecular ligation had been favoured over the intermolecular reactions (the tendency for which depends on the relative concentrations of plasmid and genomic DNA), in which case most of the plasmids in transformed cells would not be recombinant.

6. **High copy number.** Having a high copy number is desirable but not essential, like having markers for DNA insertion. To maximize the yield of plasmid from transformed cells, the copy number in each cell should be as high as possible. Different plasmids have different copy numbers. For some, such as the F factor, the copy number is low, perhaps one or two per cell. The replication of the F factor is quite closely tied to replication of the chromosome, and such control of replication is said to be stringent. The plasmid pSC101 (one of the first plasmids used for cloning) is also under stringent control and is usually present at no more than five copies per cell. Other plasmids with different origins of replication have less tightly controlled replication, and they are said to be relaxed. Here, the copy number can range widely. The copy number of pBR322 is typically 15–20 per cell, whereas vectors such as the pUC plasmids may be present in 500–700 copies per cell. The origin of replication for both pBR322 and the pUC plasmids came ultimately from the plasmid pMB1. The reason for the difference in copy number between these plasmids with, in theory, the same replication origin may lie in a mutation in the pUC plasmids in the region encoding the RNA molecules that regulate replication by interacting with the origin. This mutation makes the repression system that controls plasmid replication less effective. The copy number of many low copy-number plasmids containing the pMB1 origin can be increased by chloramphenicol amplification. The host cell culture is treated with chloramphenicol, which inhibits bacterial protein synthesis. The inhibition of protein synthesis blocks chromosomal DNA replication, because particular proteins, such as DnaA, need to be synthesized each time chromosomal replication is initiated. The inhibition also blocks cell division, which is closely tied to chromosomal DNA replication. Plasmid replication, however, requires only proteins that are more long-lived. It can, therefore, continue when chromosomal DNA replication and cell division have stopped. Eventually, plasmid DNA replication will stop too, as the supply of general replication proteins, such as DNA polymerase, runs out; but the average copy number will have increased greatly.
7. **Disablement.** Ever since the earliest experiments in genetic manipulation, there has been concern over the possibility of recombinant DNA molecules ‘escaping’ into the environment and spreading. The likelihood of this can be reduced if the plasmid is in some way disabled so that it cannot spread to other bacteria by processes such as conjugation. Many plasmids, such as pBR322,

have been disabled by removal of the *mob* gene, which is required for them to mobilize themselves by conjugation. However, such plasmids can still be transmitted from cells containing other plasmids that can provide the necessary functions for mobilization. This transmission can be blocked by removing a region containing sites called nic and *bom* from the plasmid that might be mobilized. This region is where the proteins provided by the other plasmids act. The pUC vectors are among those that have had this region removed.

3.2.2 Transformation

The transformation method chosen will depend on the efficiency required. The more complex chemical methods for inducing competence described in Section 1.5.2 can generate up to 10^9 transformants per microgram of plasmid DNA, with up to 5% of viable cells competent for transformation. If the recombinant molecules are abundant or of just one (or a few) types, as when reintroducing a plasmid DNA stock into cells to prepare more of the same plasmid, then simple treatment with calcium chloride will be sufficient. If the number of recombinant molecules available is very small and as many different members of the collection as possible need to be recovered (e.g. in cDNA cloning, see Section 5.3), then a more efficient system such as electroporation may be required. When optimized, this approach can yield a higher number of transformants per microgram of DNA than any of the chemical treatments (up to 5×10^{10} per microgram have been claimed). Typically, a field strength of the order of 15 kV cm^{-1} is used, with a decay time constant of about 5 ms. These conditions will usually result in most of the cells remaining viable, and a reasonable balance between cell killing and the induction of competence.

3.2.3 Hosts

The host is the cell in which the recombinant molecules are to be propagated. Choosing the right host is as important as choosing the right vector. Essential or desirable characters include the following:

1. **Efficient transformation.** This depends on two main features. One is the ability actually to take DNA into the cell, and this process is poorly understood. Different genotypes respond differently to different transformation systems. Some mutations appear to enhance the efficiency of transformation itself. These include the *deoR* mutation, which seems to assist particularly in the uptake of larger DNA molecules. DeoR is a transcriptional regulator with DNA-binding activity that controls expression of a set of genes involved in deoxyribonucleoside catabolism.

   The main feature determining transformation efficiency is the presence or absence of endogenous DNA-degrading systems.

Many hosts used for cloning are derived from *E. coli* strain K, which contains the K restriction–modification system encoded by the *hsdRMS* locus. The *hsdR* gene encodes an endonuclease that cleaves DNA containing the sequence -AACNNNNNNGTGC-, unless the second of the two adenine residues and the adenine residue on the other strand opposite the thymine are methylated. Many hosts, therefore, have an *hsdR* mutation (or a larger deletion) to avoid cleavage of incoming unprotected DNA. The *hsdM* gene encodes the methylase that protects against degradation so passaging of DNA through an $hsdR^{-}M^{+}$ strain can be used to allow methylation if it is subsequently necessary to propagate in an $hsdR^{+}$ strain.

There are other proteins in *E. coli* strain K that will degrade incoming DNA if it *is* methylated, belonging to the methylation-dependent restriction systems (MDRS). These are not as well understood as the Classes I–III restriction–modification systems. They include the endonucleases that are the products of the *mcrA*, *mcrB* and *mrr* loci, which will degrade DNA containing methylcytosine (*mcrA*, *mcrB*) or methyladenine (*mrr*). Using strains mutant in these loci, therefore, is desirable, particularly when cloning highly methylated DNA.

The extent of methylation of DNA in the host may also affect the efficiency with which other restriction enzymes will subsequently cleave the DNA in vitro. Methylation can be carried out by the enzymes mentioned already, and also by the products of the *dam* and *dcm* genes. The Dam protein methylates adenines at the sequence -GATC-, and Dcm methylates cytosines at the sequences -CCAGG- and -CCTGG-. Some frequently used restriction enzymes have recognition sites that overlap with these and are inhibited by methylation, so DNA prepared from strains that are wild type for these loci will not be efficiently restricted. Use of dam^{-} or dcm^{-} strains has the disadvantage, however, that they have high mutation rates. This is because newly replicated dsDNA is hemi-methylated (as the newly synthesized material has not yet been methylated). If the polymerase has introduced any errors during synthesis, they will result in a mismatch; these mismatches are normally resolved in the cell by correction to the *methylated* strand. In a dam^{-} or similar strain, neither strand is methylated, and the mismatch correction is as likely to be to the incorrect (newly synthesized) strand as to the correct one.

2. **Stable maintenance of plasmid.** Once a recombinant plasmid has entered the cell (assuming it has escaped any endogenous restriction enzyme activity), it is still not guaranteed to replicate indefinitely and stably even if it has a suitable origin of replication. Rearrangement of the recombinant may occur, and the most frequent manifestation of this is partial deletion (i.e. loss of part of the molecule). This usually occurs by recombination across directly repeated sequences, as shown in Figure 3.5. Not all the plasmid molecules in one cell need undergo the same

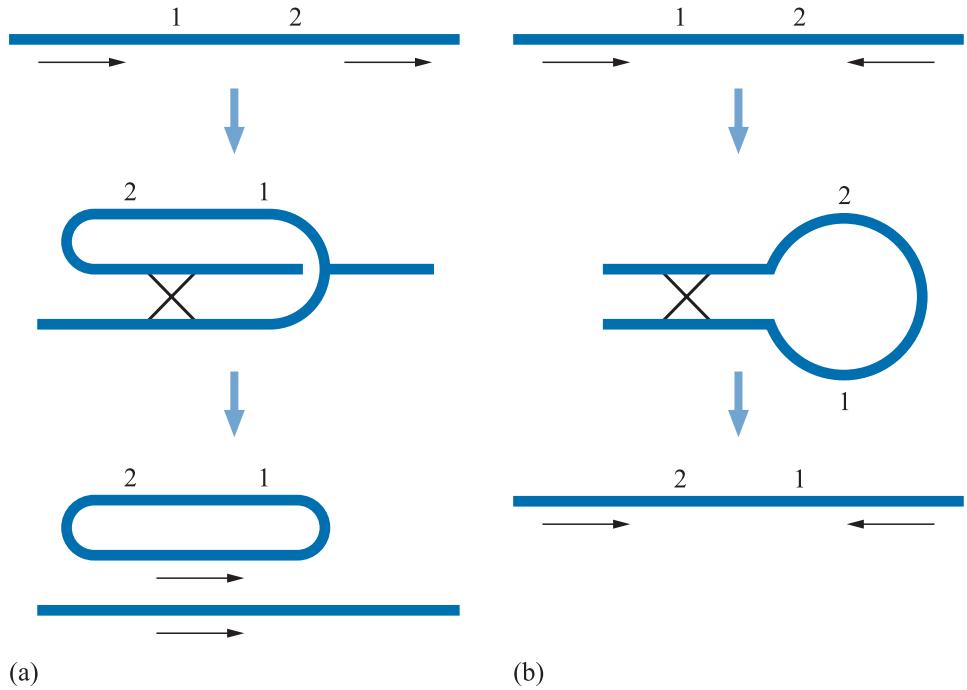


Fig 3.5 Recombination across repeated sequences. If the repeats are in a direct configuration, then the material between them is lost (a). If they are inverted, then the material between them is inverted on recombination (b). I and 2 represent arbitrary points on the molecule.

recombination at the same time, so recombination across a pair of directly repeated sequences in a plasmid will leave three types of molecule in the cell: the original, unrecombined one; and the two smaller recombination products. One of the products will lack an origin of replication and be lost from the population. The other will still be able to replicate, and will do so faster than the original molecule because it is smaller. Over a number of rounds of replication, even a small difference in the time required for replication can result in a large difference in the number of molecules present. Generating 1 μg of DNA from a single copy of a 4 kbp plasmid takes about 40 cycles of replication. A plasmid that could replicate 10% faster would, over this period, generate approximately 16 times as much DNA and, therefore, be by far the most abundant molecule in the total population. The problem will be exacerbated if the original plasmid is deleterious to the host, perhaps by expression of a protein that is toxic to the cell, and the partial deletion caused by recombination abolishes production of that protein. The possibility of preferential propagation of altered molecules during cloning is, therefore, one that should always be borne in mind. Recombination need not always result in deletion. It can also cause inversion, if it takes place across inverted repeats (Figure 3.5).

Since deletions and inversions usually depend upon recombination, mutations in the host that suppress recombination will help to ensure the stability of transforming molecules. There are three main recombination systems in *E. coli*, using the products of the *recBCD*, *recE* and *recF* genes. However, these systems depend largely on the product of the *recA* gene, so strains mutant in this will have a greatly reduced recombination frequency. Some sequences, particularly those containing inverted repeats, may be subject to rearrangement in a *recA*-independent way. This may be due to the *recF* pathway functioning in the absence of (or low levels of) *recA*, so an *recF* mutation may be desirable in addition to *recA*. It is also reported that deficiency of the DNA gyrase encoded by the *gyrA* and *gyrB* genes can lead to enhanced stability of molecules containing repeated sequences.

3. **Disablement.** As with vectors, it is necessary to take precautions to ensure that strains carrying recombinant plasmids are unlikely to escape and propagate outside the laboratory (i.e. to enhance containment). For this reason, the preferred strains usually carry mutations that reduce their viability in the wild. These are often mutations conferring auxotrophy (i.e. the requirement for a particular metabolite to be supplied in the medium, resulting from an inability to synthesize the metabolite).
4. **Features allowing use of the lacZ′ minigene**. We saw earlier that the use of the *lacZ′* minigene requires the use of a host producing a suitable LacZ fragment, and that this is sometimes encoded on an F′ plasmid. In these cases, therefore, it is necessary to be able to ensure retention of the F′ plasmid. For this reason, the F′ plasmid also contains genes *proA* and *proB* for two of the enzymes of proline biosynthesis (encoding 5-glutamyl phosphate reductase and glutamate-5-kinase respectively), and these are deleted from the chromosome of the host. Propagation of the host on medium lacking proline ensures that only cells carrying the resident F′ plasmid will grow.
5. **Other markers.** In any work with bacteria, it is very useful to have strains that are genetically marked, so the correct strains can be recognized as such. The recombination and nutritional deficiency markers mentioned above may be useful in this respect, although it may also be convenient to have other markers that can be selected more easily. Antibiotic resistance can be useful, although of course it should be different from any of the antibiotic resistances conferred by the vectors to be used. Another mutation that is often incorporated into hosts is *endA*, inactivating the gene for a DNA-specific endonuclease. This mutation enhances both the yield and quality of plasmid DNA preparations. The *relA* mutation overrides controls on RNA synthesis, improving the rate of synthesis in the absence of protein synthesis. The *tonA* mutation causes loss of an outer membrane protein, to which phages T1 and T5 adsorb. Infection by these phages can be a problem in the laboratory, so the *tonA* mutation is useful in conferring resistance

to phage infection. Host strains used for particular purposes, such as the expression of cloned genes, have additional features, as discussed in Chapter 8.

It may be helpful to look at the genotype of a common host strain, and see how it fulfils the considerations above. The strain DH5α is a common host for the pUC plasmids, and is a derivative of *E. coli* strain K. Its genotype is *endA1 hsdR17 supE44 thi-1 recA1 gyrA* (nal^R) *relA1 Δ(lacZYA-argF)U169 Φ80lacZΔM15*

endA1 is discussed above.

hsdR17 inactivates the host restriction system, allowing DNA to be protected by methylation, but not digested, which is sometimes abbreviated to $r_K^- m_K^+$.

supE44 is an 'amber' chain termination suppressor mutation that allows readthrough of UAG codons in translation. This is used because the same host (or derivatives of it) can also be used for propagation of phages, a number of which have amber chain termination mutations as a biological containment measure. Growth of such phages is possible only in a suppressor host.

thi-1 is a nutritional requirement (for thiamine) that gives some containment.

recA1 gyrA (nal^R) ***relA1*** are discussed above. (nal^R) is resistance to the DNA gyrase inhibitor nalidixic acid, conferred by the *gyrA* mutation.

Δ(lacZYA-argF)U169 is a deletion of the *lac* operon (extending into an arginine biosynthesis gene) that is required for the blue-white selection.

Φ80lacZΔM15 describes a ***Φ80lacZ*** prophage that directs synthesis of the partially deleted LacZ, required for blue–white selection.

Another example of a widely used host is JM109, also a derivative of *E. coli* strain K. It is *endA1 hsdR17 supE44 thi-1 recA1 gyrA* (nal^R) *relA1 mcrA Δ(lacZYA-proAB) F′ traD36 proA*$^+$ *proB*$^+$ *lacI*^{q} *lacZΔM15*.

Many of the mutations are similar to those in DH5α. JM109 also carries the ***mcrA*** mutation, described earlier, and a different *lac* deletion, extending into the proline biosynthesis operon. The symbol *F′* indicates that the markers following it are on an F′ plasmid.

traD36 reduces the efficiency of conjugation by a factor of 10^5, contributing to biological containment.

proA$^+$***proB***$^+$ restores the ability to synthesize proline that was lost by the deletion into *proAB*. This means that growth of the strain on a medium that lacks added proline selects for the presence of F′, as it requires the cells to synthesize their own proline.

lacI^q is a mutant form of the Lac repressor gene, causing increased levels of repressor in the cell, and allowing tighter control of the *lacZ* gene. This is important for the use of this host for vectors that replicate to a high copy number. Such vectors might titrate out all the repressor produced by a wild-type *lacI* gene, especially if the gene was present only on the chromosome or on a low-copy-number plasmid

such as F′. Titration of the repressor might lead to uncontrolled transcription of sequences inserted within the *lacZ* region of the vector.

lacZΔM15 is the partially deleted *lacZ* needed for alpha-complementation.

Notice that JM109 allows selection for retention of the F′ plasmid, as described. This is important for two reasons. The partially deleted *lacZ* needed for alpha-complementation is encoded on the F′. Also, the host is sometimes used for growing phages (Section 4.3) that propagate only in a host carrying the F (or F′) plasmid.

3.3 Modifications

3.3.1 Avoiding self-ligation

Although the blue–white screening that is available with the pUC vectors allows us to see which colonies on a plate contain recombinant plasmids, these are sometimes very few in number compared with the colonies containing vector that religated without any insert. If we need to recover a large number of different recombinant molecules, then the large number of non-recombinant colonies may be a problem. There are various approaches to minimizing this. One approach is to increase the ratio of insert DNA to vector DNA. A more reliable solution is to use alkaline phosphatase (Section 1.3.1). This enzyme is capable of removing 5′-terminal phosphate groups from nucleic acid molecules (and nucleotides). However, these phosphate groups are required for ligation to take place. So if the vector is treated with alkaline phosphatase to remove its terminal phosphate groups, then self-ligation is no longer possible. Ligation of vector to insert, however, is still possible because each insert molecule carries two phosphate groups, one at each end. Note though, that sealing each vector–insert boundary requires two phosphate groups, one for each strand. Phosphatase treatment of the vector, therefore, means that each boundary can be covalently sealed by ligase on one strand only. The other strand will remain nicked, as shown in Figure 3.6. However, these nicks are not a serious problem. Transformation is still possible, and the normal cellular replication and repair processes will produce molecules that are no longer nicked.

In practical terms, the alkaline phosphatase treatment is carried out after the vector has been cut, and before vector and insert DNA are mixed. All traces of phosphatase activity must be removed before the insert DNA is added, otherwise the insert would be dephosphorylated, too, and intermolecular ligation will also be blocked. Commonly used alkaline phosphatases are inactivated by heating for a few minutes. This treatment can be followed by removal of protein with phenol and chloroform to minimize any carry over of active enzyme into the ligation reactions. If the enzyme used is sufficiently heat labile, purification of the DNA prior to ligation can be omitted.

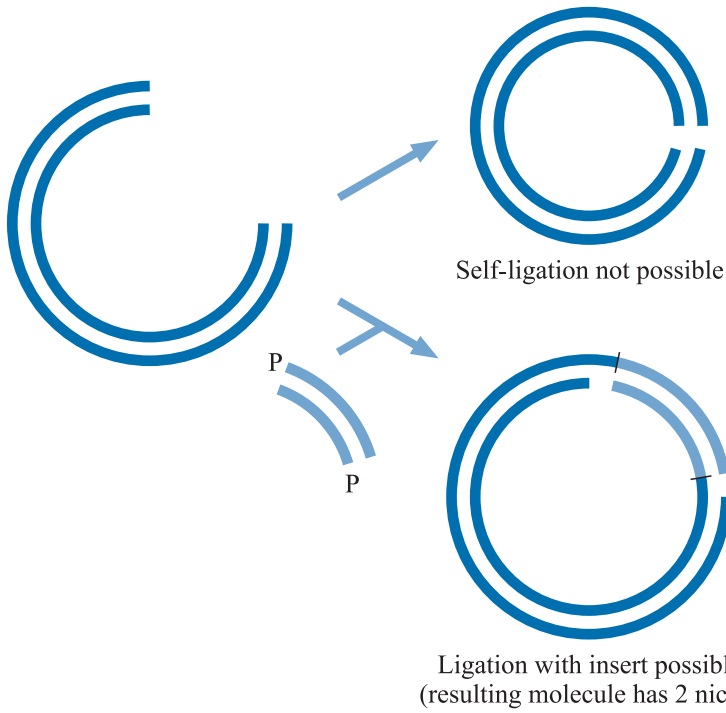


Fig 3.6 Phosphatase treatment of vector. Removal of the phosphate groups (P) from the vector means that self-ligation is not possible. Ligation to a molecule with 5′-phosphates is still possible, although one strand will be nicked at each junction.

A third approach to increasing the fraction of recombinants recovered is to use more complex vectors that, if self-ligated, direct the synthesis of a molecule that kills the host cell. One example uses the *ccdB* (control of cell death) gene. The coding region for this is arranged in the vector so that self-ligation results in the expression of the CcdB protein and consequent death of cells containing the molecule. This is the basis of 'Zero Background™' vectors.

3.3.2 Forced cloning

In the example we considered, the insert DNA could have been ligated into the vector DNA in either of two orientations. In some circumstances it is desirable to control the orientation in which a particular restriction fragment is inserted. This is sometimes called forced cloning. This is possible with the pUC vectors, as there are many different restriction sites clustered in the multiple cloning site. The region is also sometimes called the polylinker. The consequence of having this collection of sites is that, if the insert DNA has different restriction sites at each end, we can readily use a vector cut with the two corresponding enzymes. There will then be only a single orientation in which the insert can be ligated to the vector.

There is a series of pUC vectors with different combinations and arrangements of restriction sites in the multiple cloning site. They generally come in pairs whose multiple cloning sites are mirror images of one another, as shown in Figure 3.7. They were generated

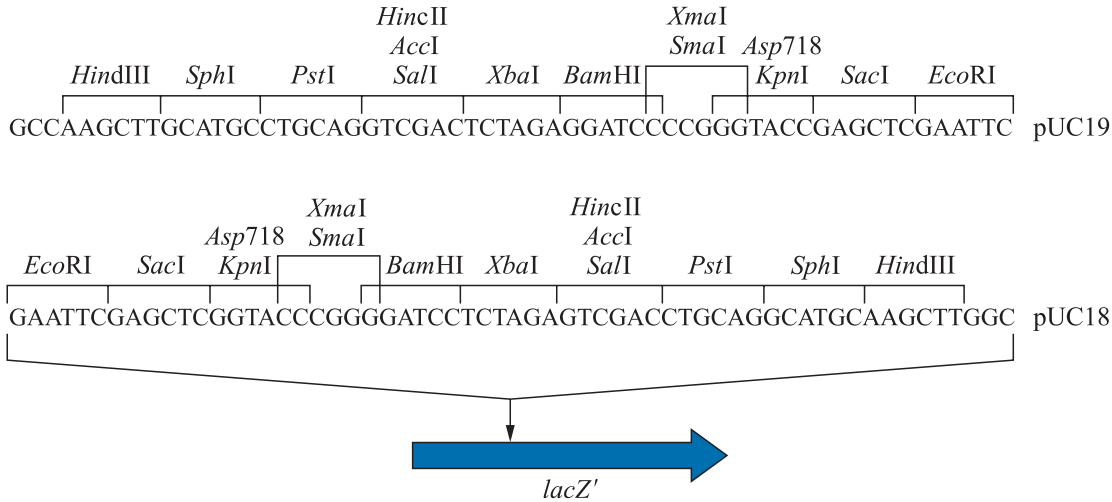


Fig 3.7 Multiple cloning sites of pUCI9 and pUCI8. Note that they contain the same restriction sites but in the opposite orientation. So an *Eco*RI—*Hin*dIII fragment, say, could be cloned into either vector, but would have opposite orientations.

by the ligation of different synthetic DNA molecules containing the appropriate restriction sites into an *Eco*RI site close to the beginning of the *lacZ′* minigene.

This wide combination of restriction sites, combined with the blue—white screening, makes the pUC vectors very versatile. There are many other series of vectors based on the same principle, but with additional features (e.g. to drive transcription of the insert DNA). We shall deal with some examples of these later.

3.3.3 Cloning PCR products

The basic cloning method can be applied to the cloning of PCR products. In the simplest case, we can treat the PCR products as blunt-ended DNA molecules and ligate them into a vector that has been cut to give blunt ends. Very often, this blunt-end ligation is not very efficient. A better approach is to incorporate restriction sites into the primers used for PCR. The PCR products will, therefore, have these restriction sites located at each end and they can be cut with the appropriate enzyme(s) prior to a sticky-ended ligation into an appropriately cut vector. It can be even more efficient to exploit the fact that many polymerase preparations used for PCR incorporate an additional A residue that is not template encoded onto the end of the molecules they synthesize. The products can, therefore, be ligated directly into vector molecules that have an overhanging T residue. A suitable vector DNA preparation can be made by cutting vector with an enzyme that generates blunt ends and then enzymatically adding a 3′ T residue to the ends. This is the basis of the pGEM-T vector series, which are available commercially as linearized plasmids with the T residue added.

3.3.4 Methods that do not use DNA ligase

In the basic experiment, we used restriction enzymes to cut DNA molecules and ligase to rejoin them. There are other systems available that have different ways of cutting and ligating molecules. They are often faster than approaches based on DNA ligase, but they are newer and less widely used.

1. **Topoisomerase.** This approach uses the *Vaccinia* virus topoisomerase I. In vivo, it cleaves one strand of a DNA molecule at the recognition sequence -C/TCCTT- and becomes covalently attached to one end of the cleaved molecule. The supercoiling of the DNA template can then be altered by swivelling around the uncleaved strand, and the enzyme then reseals the original strand and dissociates from the DNA. Commercially available topoisomerase-based vectors ('TOPO® Cloning' systems) are supplied in linear form, with the topoisomerase enzyme attached. The topoisomerase will rapidly ligate the vector to suitable substrate molecules. For cloning PCR products with overhanging A residues, the vector has a single overhanging T residue with the topoisomerase attached. The PCR product can anneal to the vector and similarly be rapidly ligated.
2. **Recombinase.** Site-specific recombination reactions can be used to clone PCR products or transfer inserted pieces of DNA rapidly from one vector to another. This is the basis of the 'Gateway®' system, based on bacteriophage lambda (Section 4.4). This bacteriophage can integrate its DNA by site-specific recombination between a site on the phage (*attP*) and a site on the host genome (*attB*). Recombination across *att* sites is exploited in the Gateway system to transfer inserts directly between vectors carrying *att* sites. For cloning PCR products, for example, the primers contain the *attB* site. The recipient vector (called an entry vector) contains *attP* sequences flanking the insertion site. Addition of recombinase protein catalyses recombination between *attB* and *attP* sites, integrating the PCR product into the vector, as shown in Figure 3.8. The insert can then be transferred, if required, into different vectors that have suitable *att* sites by similar site-specific recombination. Alternatively, restriction fragments can be inserted into an entry vector by conventional ligation and then transferred into other vectors by recombination.

3.4 Linkers, adaptors and cassettes

3.4.1 Linkers

The importance of the pUC and similar vectors owes a great deal to the presence of the multiple cloning site, with its ability to accept a wide range of different fragments. A similar feature is offered by

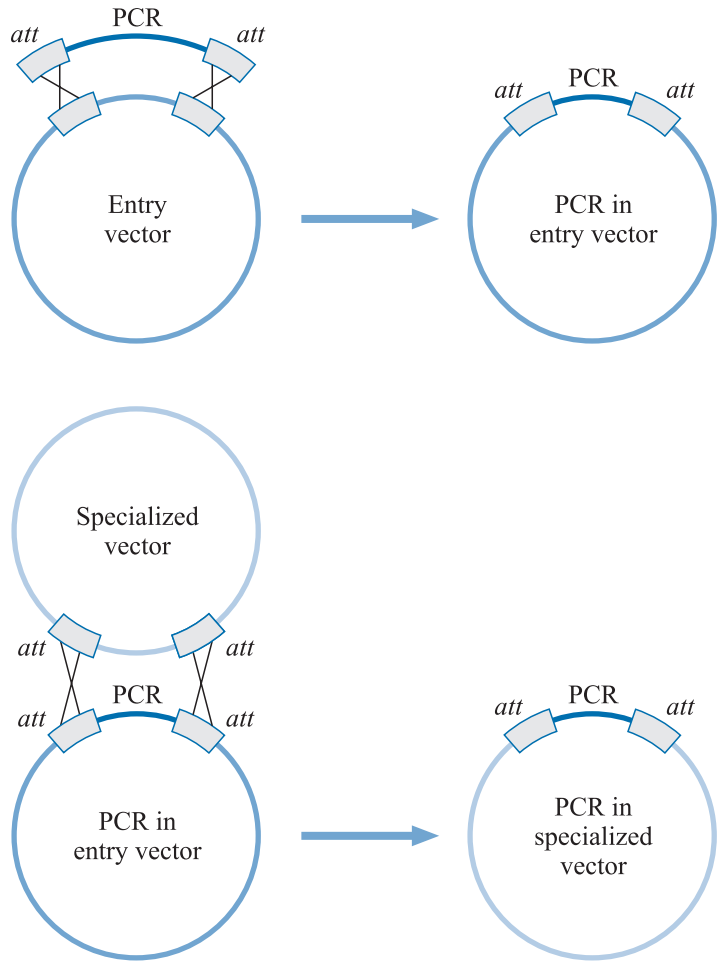


Fig 3.8 'GATEWAY®' cloning system. In the first stage a PCR product (PCR) is transferred into an entry vector by site-specific recombination across *att* sites. In the second stage, the product is transferred into a specialized vector by recombination.

molecules called linkers. These are short, chemically synthesized molecules that contain a particular restriction enzyme recognition site within their sequence. An example of these, an *Eco*RI linker, is shown in Figure 3.9. Using *Eco*RI linkers, it is possible to clone a blunt-ended insert molecule into a vector carrying an *Eco*RI cloning site. The linkers are themselves blunt-ended molecules, and can be joined onto the blunt-ended insert molecule using T4 DNA ligase. Although the reaction (being a blunt-ended ligation) is relatively inefficient, the use of an excess of linker helps to ensure that a large proportion of the insert molecules have linkers on the ends. Some molecules end up with more than one linker attached to each end, but that is not a problem. The next step is to treat the insert-linker molecules with the appropriate restriction enzyme, which cuts within the linkers to leave a single cut linker attached to each end of the insert. So the insert now has sticky ends, which can be used for insertion into a restriction site in the usual way. This is summarized in Figure 3.9.

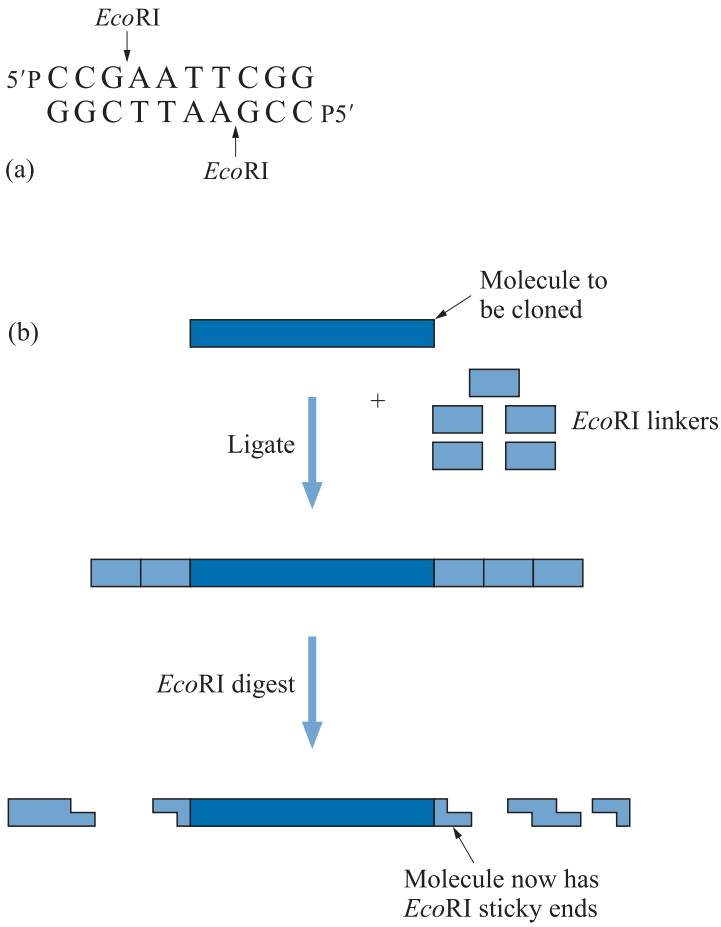


Fig 3.9 An *Eco*RI linker. An example of a linker is shown in (a). The use of linkers is shown in (b). The linkers are first ligated onto a target molecule in a blunt-ended reaction and are then cleaved with *Eco*RI.

There is a potential problem, though. If the insert itself contains a recognition site for the linker restriction enzyme, then the second step of the process in Figure 3.9 will cut the insert as well as the linkers. This is likely to be undesirable. It can be avoided by treatment of the insert, prior to the addition of linkers, with the appropriate methylase: *Eco*RI methylase if *Eco*RI linkers are being used. This renders the molecule insensitive to the restriction enzyme, which will still be able to cut the linker.

Of course, an alternative way of cloning blunt-ended DNA into a vector with *Eco*RI ends is to avoid using linkers and instead polish up the ends of the cut vector to make them blunt. That could be followed by a blunt-ended ligation of insert DNA into vector. What are the advantages of using linkers? There are three main advantages. One is that it makes better use of the insert and vector DNA, which may be in short supply. Ligation of blunt-ended insert into blunt-ended vector is an inefficient reaction, with much of the insert being wasted. Ligation of blunt-ended linker onto blunt-ended insert can also be inefficient. However, we use a large excess of linkers

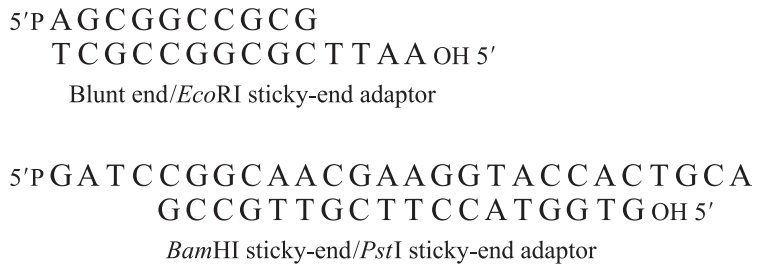


Fig 3.10 Adaptors. The upper molecule has a blunt end (left-hand side) and an *Eco*RI sticky end. It can be used to convert a blunt-ended molecule into an *Eco*RI-ended molecule. The lower molecule has a *Bam*HI end (left-hand side) and a *Pst*I end (right-hand side). Therefore, it can be ligated onto a molecule with *Bam*HI ends to produce one that has *Pst*I ends (or vice versa, depending on the availability of phosphates on the target molecule).

(which are in abundant supply, being synthesized on a chemical scale), so that a large proportion of insert acquires linkers and can be efficiently ligated into the vector. The second advantage, which also comes with using an excess of linkers, is that the likelihood of two insert DNA molecules being ligated to each other (concatenated) is reduced. The third advantage comes when it is necessary to cut the insert out of the vector at a later stage. It is unlikely that ligation of a blunt-ended insert into vector that has been cut with *Eco*RI and then polished will regenerate any restriction sites. It is difficult, therefore, to excise the insert precisely from the vector. However, insert cloned into *Eco*RI-cut vector using *Eco*RI linkers can readily be excised simply by redigestion with *Eco*RI.

There is no reason why a linker should have only one restriction site within it. Linkers that contain a number of sites are available. These are often called polylinkers, and resemble the multiple cloning site of the pUC vectors.

3.4.2 Adaptors

One or both ends of a linker may be single stranded. These are sometimes called adaptors. Examples are given in Figure 3.10. The first shown is an adaptor that is blunt at one end (like a conventional linker) and sticky at the other. Therefore, it can be ligated, without further digestion, to a blunt-ended molecule to leave a sticky end. Note that in this example the 5′ overhanging end is not phosphorylated; this lack of a phosphate group prevents the concatenation of adaptors in the ligation reaction (which would obscure the sticky ends to be used for cloning). Some adaptors (such as the second example in Figure 3.10) are sticky at both ends and can be attached to a molecule that is already sticky-ended. Adaptors may also have extra restriction sites within their sequence.

3.4.3 Cassettes (cloning cartridges)

Cassettes (cloning cartridges) are a combination of linkers with other features, such as selectable markers. In their simplest form they consist of an antibiotic resistance gene flanked by DNA that contains multiple cloning sites (Figure 3.11). Consequently, they can be used as an easy way of incorporating selectable markers or other features into DNA molecules. This can be particularly useful as a way of inactivating a cloned gene in the technique of gene disruption

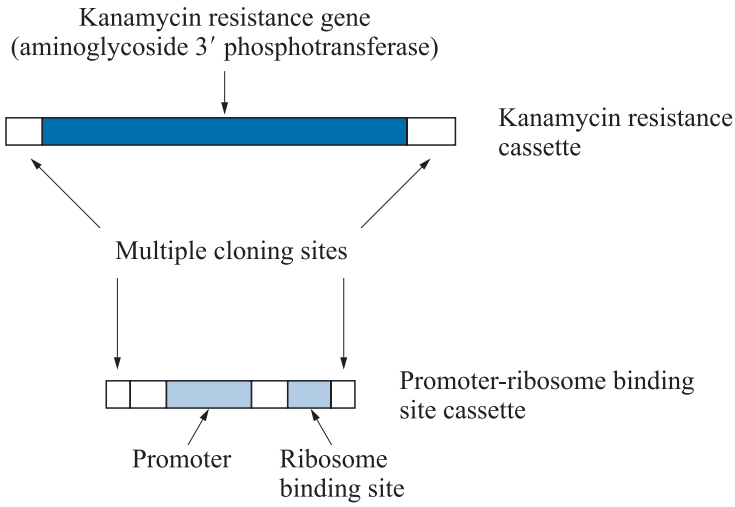


Fig 3.11 Cassettes. The diagram shows examples of an antibiotic resistance cassette (upper) and an expression cassette (lower).

(see Section 7.6). Cassettes may contain gene expression signals (such as promoters, terminators, etc.) rather than selectable markers. These cassettes are called promoter cassettes, terminator cassettes, and so on.

Chapter 4

Other vector systems for *E. coli*

4.1 Introduction

So far, we have considered the use of small plasmids as cloning vectors for *E. coli*. However, these are not the only molecules able to replicate inside bacterial cells. For example, the *E. coli* F (fertility) factor is a large plasmid that can replicate independently of the bacterial chromosome, like other plasmids (although it can also insert itself into the bacterial chromosome). The F factor is used as the basis of bacterial artificial chromosome (BAC) vectors, for cloning very large pieces of DNA. Bacteriophage viruses are also able to replicate inside the bacterial cell, and a number of them have been developed for use as cloning vectors. These include the phages M13, lambda, Mu and P1. The M13 vectors have a lot in common with the pUC vectors we looked at in Chapter 3. They can be used for generating single-stranded DNA and they have been particularly useful in DNA sequence determination, although they are now less widely used for this. The lambda vectors are used for more general cloning purposes. Phage Mu is often exploited for its ability to act as a transposable genetic element, and Phage P1 is particularly useful for the production of phage artificial chromosomes (PACs). Like BACs, these can be exploited for the cloning of very large pieces of DNA. There are also a selection of vectors that are hybrids between plasmids and phages.

4.2 BAC vectors

BAC vectors are based on the F (fertility) factor that was identified in *E. coli* as catalysing genetic exchange between cells. The F factor is a plasmid that can direct its own transfer into cells that do not already contain one. Cells containing an F plasmid are sometimes referred to as male cells. The transfer is mediated by a proteinaceous filament, the sex pilus. As well as existing as a separate molecule, the F factor can integrate into the host chromosome to form an

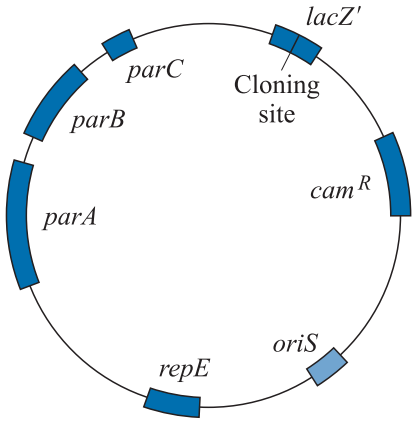


Fig 4.1 The BAC vector pBeloBAC11 (7.4 kb). The markers and other functional sequences are described in the text.
For simplicity, *cos* and *lox* sites that are also present have been omitted.

Hfr (high frequency of recombination) strain. Aberrant excision from the host chromosome in vivo can generate forms of the F plasmid that contain additional pieces of chromosomal DNA. These are called F′ plasmids. Because the F plasmid is able to accommodate additional DNA sequences, which can be very large, it can also be used as the basis of cloning vectors able to propagate large DNA inserts. The F plasmid is usually present at low copy number, typically one or two molecules per cell. The low copy number increases the stability of the molecule, since there are fewer copies to act as potential substrates for recombination-mediated deletion. This is helpful when the BAC vector is used to clone very large pieces of DNA, when instability would be a problem. The disadvantage of this is that the yield of DNA on extraction from cells is low – but the increased stability when cloning large pieces of DNA usually outweighs this. BAC vectors can conveniently be used to accept inserts of 100–300 kbp. An example of a BAC vector is shown in Figure 4.1. The *repE* and *oriS* sequences are required for replication, and the *parA–C* sequences regulate the copy number. There is a selectable marker, for chloramphenicol resistance (cam^R), and a *lacZ′* sequence containing the cloning site and allowing screening for the presence of an insert. Recombinant BAC vectors can be conveniently introduced into *E. coli* by electroporation.

4.3 Bacteriophage M13 vectors

4.3.1 General biology

Bacteriophage M13 is particularly useful because of its slightly unusual life cycle (summarized in Figure 4.2), which gives us a way of getting single-stranded DNA from dsDNA. It belongs to a group called the filamentous or 'skinny' phages, on account of their dimensions. A typical example of the filamentous phages, fd, has dimensions 850 nm × 6 nm × 6 nm. The infectious particle contains single-stranded DNA contained within a protein coat made up primarily of subunits of a single protein species that is the product of the

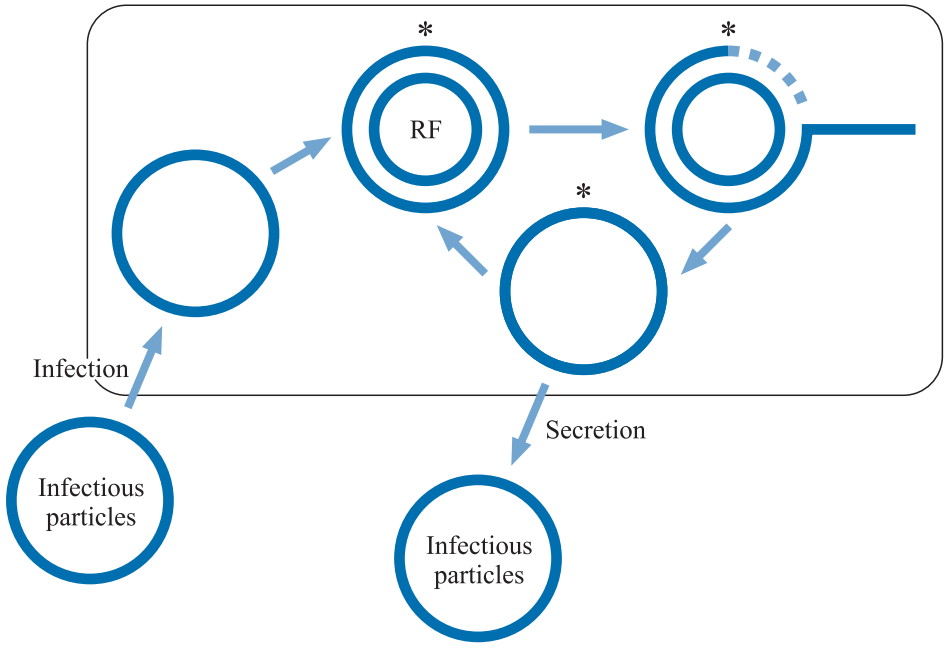


Fig 4.2 Life cycle of bacteriophage MI3. RF = replicative form, *indicates the site of nicking or closing by gene II protein and the dashed region indicates material newly synthesized by rolling circle replication.

phage gene VIII. There are also a few molecules of the product of the phage gene III. The phage gains entry to the bacterial cells by attaching to the sex pilus before passing its DNA into the cell. Therefore, it will infect only male cells (i.e. those carrying an F or F′ factor, or Hfr strains).

Once inside the cell, the single-stranded molecule (referred to as the plus strand, +) is converted to a double-stranded molecule called the replicative form (RF). This is done by a means similar to the normal replication process of *E. coli*, using a specific origin of DNA synthesis on the single-stranded molecule for synthesis of the complementary minus strand. The minus strand can then be transcribed to produce viral proteins. The RF DNA can also be replicated by a rolling circle mode of replication. For rolling circle replication, the product of gene II binds to a specific site on the double-stranded genome and creates a nick in the +strand, generating a free 3′-hydroxyl. This strand is extended by DNA polymerase, displacing the original + strand. After a round of synthesis, the displaced +strand can be separated from the newly synthesized one by another nick by gene II protein. This produces a separate +strand, and it is closed by the gene II protein to yield a circular molecule. This can again be converted to the RF, which can accumulate to 100 or so copies per cell. As the RF accumulates, so does another phage protein, the product of gene V. This blocks not only the synthesis of gene II protein (reducing the synthesis of single strands), but also the conversion of + strands to the RF. The single strands are encapsidated in coat proteins and leave the cell. Unlike most phages, virions can be secreted from the cell without causing lysis. A useful feature is that there is no clear constraint on the size of DNA molecule that can be packaged by the coat protein. If DNA molecules longer than normal are to be

packaged, then more coat protein is attached. This contrasts with phage lambda, for example, which has clear constraints on the amount of phage DNA that can be packaged into each phage particle. Importantly, bacteriophage M13 offers us a way of generating single-stranded recombinant DNA. Restriction endonucleases will work only on double stranded molecules, and the plasmids we have encountered so far are also double stranded; the best way of obtaining cloned single-stranded DNA would be to melt the two strands apart and use them before they reannealed. However, if we can construct double-stranded molecules using M13 RF as a vector and introduce them into *E. coli* by transformation, then they will behave there like normal RF DNA and generate single-stranded copies of the same molecules, which would be packaged and released from the cell. So M13 offers a way of converting dsDNA into single-stranded DNA. It is usually much easier and more reliable to use M13 than to prepare single-stranded DNA from double-stranded plasmids by denaturing them.

4.3.2 Design of M13 vectors

The same principles that we saw in Chapter 3 applied to the design of plasmid vectors also apply to the design of M13 vectors. We need an origin of replication. The phage has its own, so that presents no difficulty. There is also no problem over a selectable marker, as the phage is, in a sense, a selectable marker itself. Cells that take up phage DNA will produce more phage and become a focus for infection of other cells. Plating out the cells from a transformation will, therefore, generate a lawn (a uniform layer of cells) peppered with holes (plaques) arising from infected cells that have taken up phage DNA molecules and produced more phage, which have infected surrounding cells. Although M13 does not actually lyse the cells it infects, it retards their growth; so the plaques that are seen are not true plaques (which arise from cell lysis), but are areas of slow growth. The cell division time of infected *E. coli* increases from 20 min under optimum conditions to more than 2 h upon M13 infection. Because the plaques result from slow growth rather than genuine lysis, they are sometimes known as pseudoplaques.

The ability of phage to infect cells, produce more phage and, consequently, produce a plaque depends on inserting DNA without affecting any of the phage genes necessary for replication. The M13 genome is quite tightly packed (Figure 4.3); therefore, there are rather few places available to insert DNA. One available, site is in an intergenic spacer between genes II and IV. The *lacZ′* minigene and multiple cloning region (discussed in Chapter 3) have been incorporated here. This modification brings with it all the advantages of the multiple cloning site (such as the ability to accept a wide range of restriction fragments and to force cloning in a particular orientation) as well as the added advantage of direct colour screening for the presence of an insert afforded by the *lacZ* cloning system.

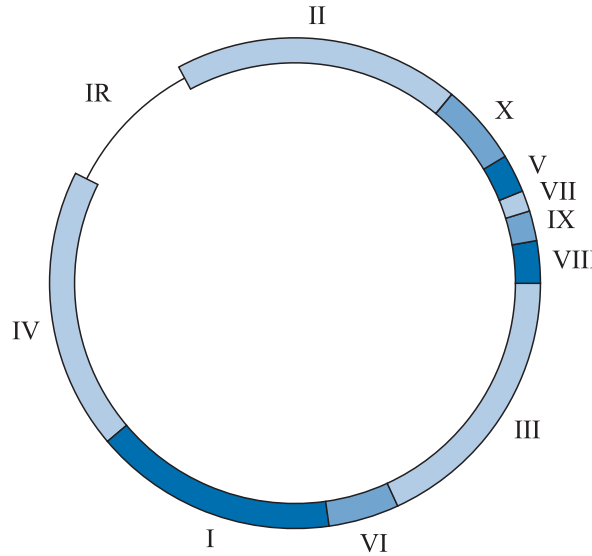


Fig 4.3 Bacteriophage MI3 genome. I—X indicate the genes; IR indicates the intergenic region where DNA can be inserted.

For the *lacZ* system to work, the rest of the beta-galactosidase gene must be present in the host. (As we saw in Chapter 3, this is often present on an F′ plasmid. An F′ plasmid is also needed to ensure that the host produces the sex pilus required for M13 phage infection.) After transformation, the host cells are mixed with X-gal and IPTG before plating out and incubation to form a lawn. Cells that have taken up phage DNA will give rise to plaques, and if the *lacZ′* coding sequence within the phage is intact, then the plaques will have a blue colour distinguishable from the off-white background. Figure 4.4 shows the organization of some of the M13 phages used for cloning. Just as with the pUC plasmids, there is a series of phages, designated 'mp'. The M13mp18 cloning region is identical to that in pUC18, and so on. The requirements for M13 hosts are similar to those used for pUC plasmids (indeed the same strains are generally used), which were discussed in the Chapter 3. Other regions for the incorporation of DNA, e.g. between genes VIII and III, have also been exploited, but these vectors are less widely used.

Other requirements for vectors are also satisfied by M13; at 7.3 kb, the vector is small enough to be handled easily, and the RF has a high copy number inside the cell. When using M13 to make single-stranded DNA, the intracellular copy number is less important than with plasmid vectors, as we are usually interested in the single-stranded DNA in the phage secreted from infected cells. Whereas intracellular plasmid DNA is collected by cell lysis as described in Chapter 1 (and this is the approach taken for preparing RF DNA for cloning purposes), the single-stranded DNA must be collected rather differently. An infected culture is set up and after a while the cells are removed by centrifugation, leaving the phage in the supernatant. The phage are then precipitated, e.g. by the addition of polyethylene

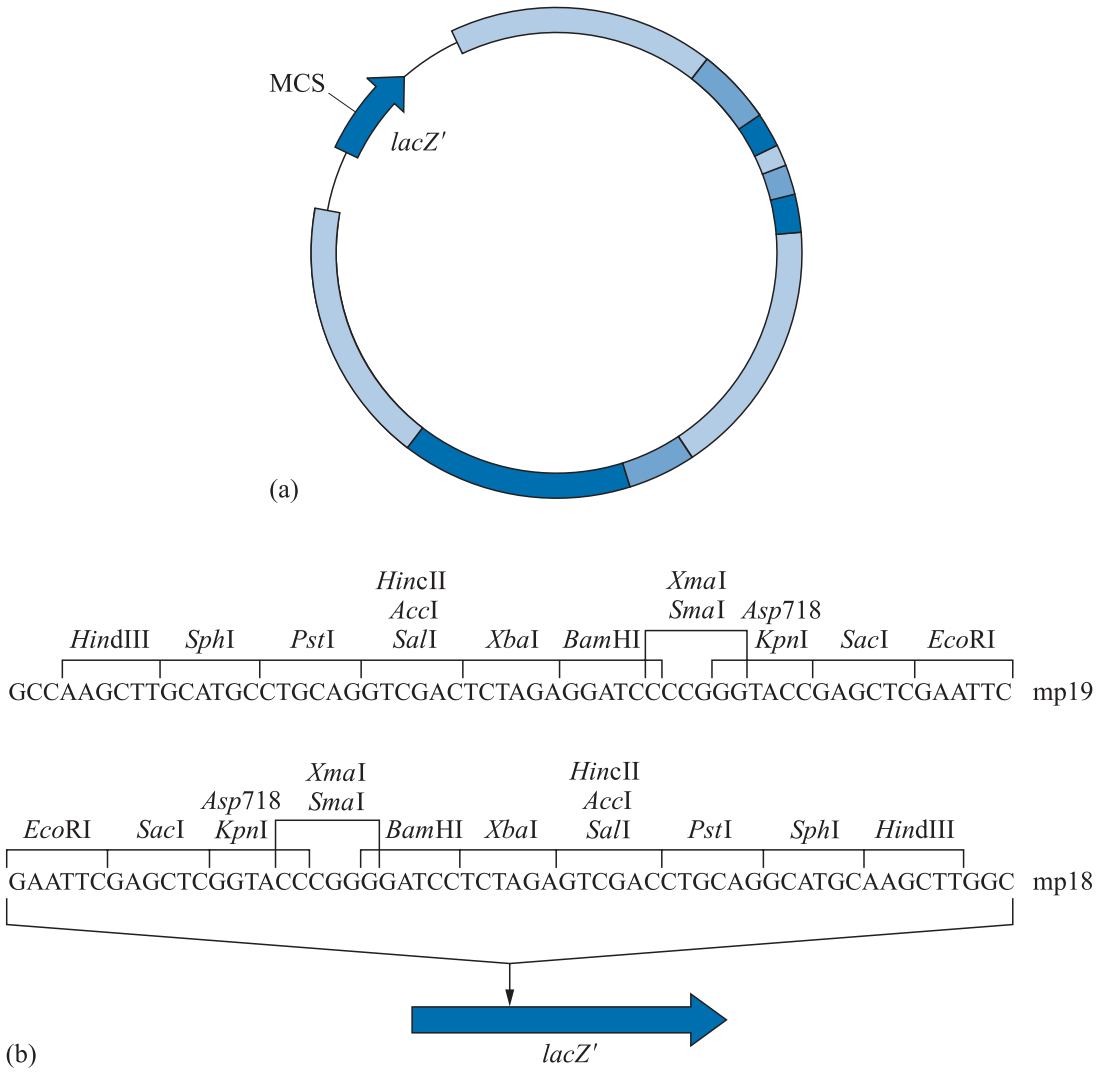


Fig 4.4 Bacteriophage M13 vectors. (a) The position of the *lacZ′* minigene and multiple cloning site (MCS). (b) The position of restriction sites within the multiple cloning sites of phages M13mp19 and M13mp18.

glycol and sodium chloride (although other methods are also possible), followed by centrifugation. The phage pellet is resuspended and protein removed with phenol, to leave single-stranded DNA in solution. This is then recovered by precipitation. Under standard conditions, several micrograms of single-stranded DNA can be obtained from a few millilitres of a culture of infected cells.

4.3.3 Uses of bacteriophage M13

1. **DNA sequencing**. For a long time the most important application of M13 cloning was in DNA sequence determination by the Sanger method, also called the dideoxy or chain-termination method. This relies on synthesis of DNA in the presence of chain-terminating inhibitors, the 2′,3′-dideoxynucleoside triphosphates (ddNTPs), shown in Figure 4.5. The method is now a very standard tool of molecular biology and will not be discussed in detail here.

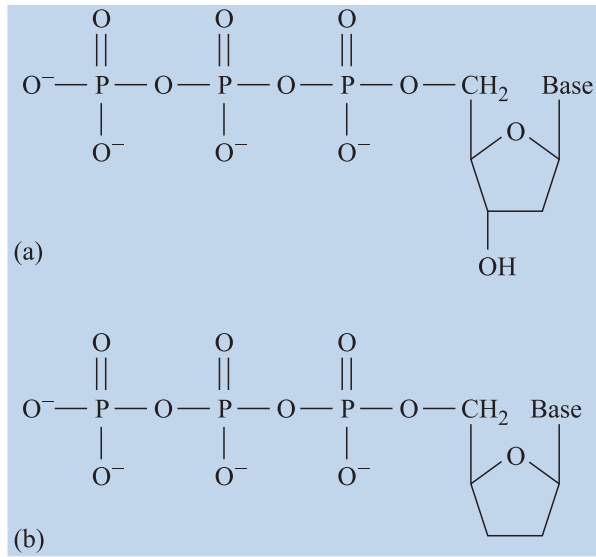


Fig 4.5 2′-Deoxynucleoside triphosphate (a) and 2′,3′-dideoxynucleoside triphosphate (b).

As initially developed, the method relied on single-stranded DNA as a template; therefore, bacteriophage M13 was routinely used for a long time as a vector to provide such single-stranded DNA for sequencing work. However, with improvements in technology, sequencing of double-stranded material is now routine, avoiding the need for the additional step of cloning into M13 vectors.

2. **Phage display vectors.** An important use of filamentous phage is in phage display systems. Here, coding sequences are inserted into one of the coat protein genes, often gene III. The result is that the phage are generated with a hybrid form of this protein, which is a fusion of the normal protein sequence and the protein product of the inserted sequence (assuming the inserted sequence has the same reading frame as gene III). The phage are secreted from the cell, with this extra material 'displayed' on the outside. Fusions with gene VIII can also be engineered, resulting in the display of many more copies of the extra peptide than with gene III, as the former is present in more copies per phage particle. These display vectors have many uses, e.g. in screening libraries by panning, as described in Chapter 6, and for vaccine production.
3. **Other applications.** Some protocols for site-directed mutagenesis (discussed in Chapter 7) also use single-stranded DNA, which can be obtained with vectors based on filamentous phages. Single-stranded DNA is also of particular use in generating probes for RNA analysis. Probes can be prepared that are specific for RNA transcripts from either strand of DNA. The latter applications are outside the scope of this book, but more information can be obtained from specialized laboratory manuals.

4.3.4 M13 derivatives

The requirements for a molecule to be replicated as single-stranded DNA and packaged into a phage coat are modest. All that is necessary is for the molecule to contain the origin of viral DNA synthesis, as long as the other functions can be provided by other DNA molecules within the cell, acting in *trans*. This fact allowed the construction of M13–plasmid hybrid vectors, sometimes called phagemid, phasmid or plage (although the term phasmid is often used for lambda phage derivatives). Plasmids that carry the M13 replication origin in addition to a conventional origin of dsDNA synthesis can be replicated either as dsDNA from the latter or as single-stranded DNA from the M13 origin. Replication from the M13 origin requires the appropriate proteins (such as gene II protein) to be provided from a helper phage also replicating within the cell. Replication generates single-stranded DNA which can then be packaged into phage coats. Examples of phagemids are the vectors pUC118, 119 and 120. They are replicated as plasmids until the cell containing them is co-infected with a helper phage, such as M13KO7, which provides the proteins for single-stranded DNA synthesis and packaging. M13KO7 is an M13 phage that has been modified, most importantly by the incorporation of a plasmid replication origin. Replication from this origin allows the helper phage to be present in a high copy number per cell and, therefore, to provide the larger quantities of the proteins that are required to replicate and package the phagemid molecule. M13KO7 also contains a kanamycin resistance gene to allow for selection for the presence of the helper phage. (Of course, it is possible that the M13KO7 helper phage may be packaged too, but in practice the packaged phagemid molecules are found to be in a 100-fold excess over the helper phage.) Another example of these vectors is the pBluescript series, such as pBluescriptIIKS$^+$, shown in Figure 4.6. This series of plasmids contains, in addition to features already described, promoters from the *E. coli* bacteriophages T3 or T7, which are useful for expressing

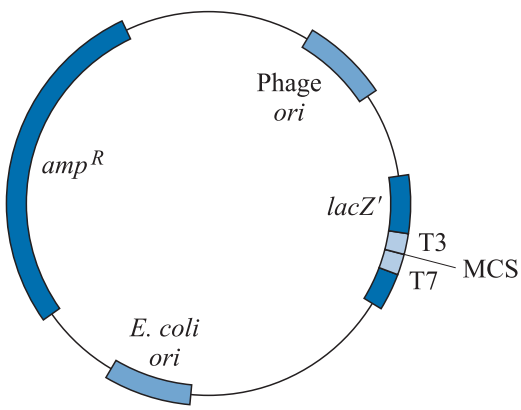


Fig 4.6 pBluescriptIIKS$^+$ (3.0 kb). The phagemid contains an ampicillin resistance selectable marker (*amp*^{R}) and origins (*ori*) for double-stranded replication (*E. coli*) and for single-stranded DNA synthesis (Phage), the latter for use when cells containing the vector are co-infected with a suitable helper phage. There are a multiple cloning site (MCS) in the *lacZ*′ minigene (allowing blue–white selection for the presence of an insert) and phage T3 and T7 promoters for transcription of inserted DNA sequences.

cloned sequences (see Chapter 8). The main advantage of the phagemid system is that it can be used to provide single- or double-stranded material without any re-cloning.

4.4 Bacteriophage lambda

4.4.1 General biology

In order to understand the cloning vectors based on bacteriophage lambda, some general knowledge of lambda biology is necessary. A simplified map of the lambda genome is shown in Figure 4.7. Lambda is an example of a temperate phage. After infecting a host cell, temperate phage can either replicate and cause lysis (the lytic pathway) or integrate their genome into the host cells to generate a lysogen (the lysogenic pathway). At a later stage the lysogen may be activated, or induced. The phage genome is then expressed and excised from the host genome, more phage are produced and the host is lysed. Which pathway is followed and under what circumstances a lysogen is induced depend on the physiological state of the host cell.

The infectious particle contains a linear dsDNA genome, with sticky ends. These ends are called *cos* (cohesive) sequences, and are generated by staggered cleavage of lambda DNA at cos sites during phage packaging. Upon infection, the molecule circularizes by annealing of the *cos* sites. Two promoters also become active (Figure 4.7). These are P_R and P_L, for rightward and leftward transcription respectively. They give rise to the immediate early transcripts, which terminate at rightward and leftward terminators, t_R and t_L. These transcripts direct synthesis of the *N* and *cro* gene products. The N protein allows the t_R and t_L terminators to be overridden, so subsequent transcription extends through into the

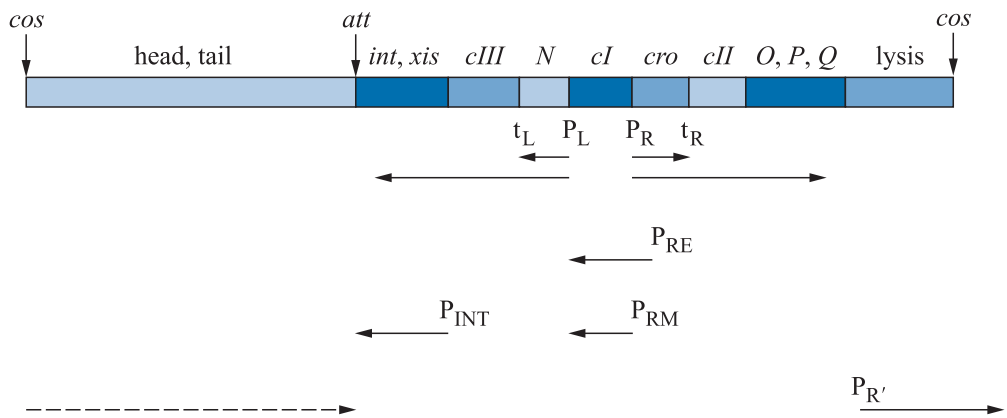


Fig 4.7 Simplified map of the bacteriophage lambda genome, with promoters and the transcripts arising from them. The molecule is shown in linear form for simplicity. Details of the promoters and transcripts are given in the text.

surrounding regions, including *cII* and *cIII*. The extended transcripts are termed the *delayed early* transcripts and include those for the proteins of DNA replication. The *cII* and *cIII* gene products, but particularly *cII*, activate another promoter, P_{RE} (promoter for repressor establishment), and it is now that the choice is made between following the lytic or the lysogenic mode.

1. **The lysogenic mode.** A major factor in the decision between lysis and lysogeny is the stability of the *cII* gene product, which is protected by the CIII protein against degradation by host proteases. (Levels of host proteases, in turn, depend upon a range of parameters, including cyclic adenosine monophosphate (cAMP) levels.) If the *cII* gene product is stable, then CII-stimulated transcription from P_{RE} leads to the production of the CI protein. This is a repressor molecule that inactivates P_R and P_L, therefore switching off the expression of the rest of the lambda genome. The CII protein and its transcript also interfere with the synthesis of Cro and Q (both of which are needed for lytic growth) and thus add a further block to the lytic pathway. The CII protein *also* activates transcription from another promoter (called P_{INT}), which leads to production of the Int protein. This protein, in conjunction with host proteins, mediates a site-specific recombination event across a site in the phage (called *att*) and a similar sequence in the bacterial chromosome, integrating the phage. However, CI repressor protein soon switches off the expression of *cII* and *cIII*, so *int* expression ceases too. The phage is thus integrated by a brief pulse of *int* expression, and phage genes are switched off by CI. The integrated lambda genome would now be completely silent transcriptionally were it not for another promoter, P_{RM} (promoter for repressor maintenance), which is activated by low levels of the CI protein (and inhibited by high levels) and leads to synthesis of further CI protein, ensuring that the repressed state of the rest of the phage genome is maintained. The integrated phage is called a prophage, and the bacterial cell containing it is a lysogen. Lysogens are immune to further infection by the same phage, because the incoming genome is immediately repressed by the levels of CI protein already present in the host cell cytoplasm. Plaques of lambda on an *E. coli* lawn will, therefore, usually be turbid, because of the growth of a few lysogens in the presence of the lambda phage in the plaque. Mutant phage that are unable to form lysogens (e.g. as a result of inactivation of the *cI*, *cII* or *cIII* genes) will, therefore, form clear plaques. (In fact, the '*c*' in those genes stands for clear plaques.)
2. **The lytic cycle.** If the CII protein is not stable under the conditions prevailing at the time of infection, P_{RE} is not activated, little repressor can be synthesized and transcription of the rest of the phage genome can continue. The Cro protein that is produced reinforces this switch by inactivating the P_{RM} promoter, ensuring that no CI repressor is produced. The Q protein that is the product

of delayed early gene expression then acts to allow expression from $P_{R'}$ (by a method analogous to that used by N), and the late genes are expressed, to produce coat (head and tail) proteins, allowing assembly of functional phage and cell lysis.

3. **Induction of a lysogen.** Induction of a lysogen will take place if the level of the CI product in the cell falls, perhaps by specific proteolysis under the action of the host RecA protein as a response to DNA damage. The fall in the level of CI lifts the repression of P_R and P_L, allowing expression of phage genes. These include *int* and *xis*, which lead to phage excision. Replication and coat protein synthesis take place, as with the lytic cycle; phage particles are assembled and the host is lysed.
4. **Replication and packaging of DNA.** Lambda DNA replication usually takes place in two phases. At first, the circular molecules generated on infection by annealing at the cos sites replicate in the bidirectional theta mode to generate additional circular DNA molecules. In the second phase, replication transfers to a rolling circle mode with a single replication fork and yields concatemeric molecules (Figure 4.8). These are needed for assembly into mature phage particles. Assembly is dependent upon the presence on the concatenated molecule of *cos* sites (Figure 4.7), flanking the region to be packaged. One genome's length of phage DNA is inserted into the phage head, and this brings adjacent *cos* sites together. Staggered cleavage of the DNA then takes place at the *cos* sites and generates 12-nucleotide overhanging ends. Without the *cos* site, packaging would not be possible. The size of the phage head imposes constraints on the amount of DNA that can be packaged into each phage particle, setting both a maximum and a minimum limit.

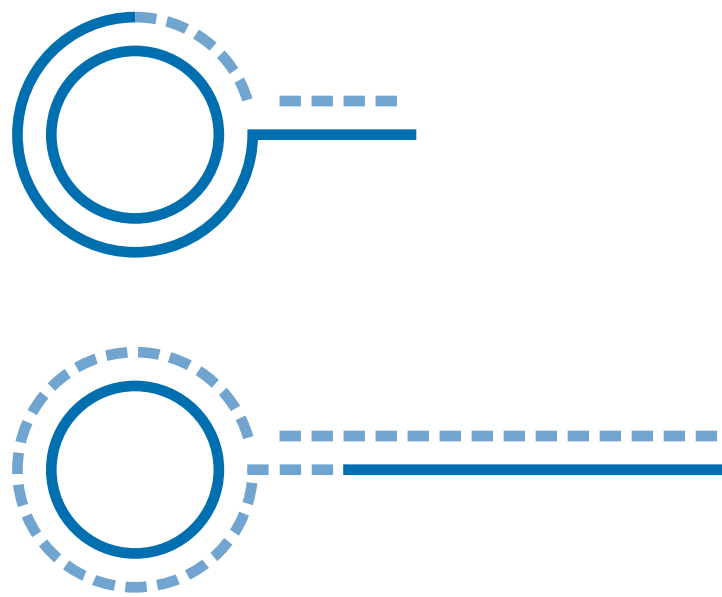


Fig 4.8 Rolling circle replication. DNA synthesis (dashed line) displaces a single strand that can also serve as a template for synthesis. As synthesis progresses (lower panel), concatenated complete double-stranded genomes are spooled off.

The wild-type lambda genome is 48.5 kbp, and molecules of between approximately 40 and 52 kbp can be efficiently packaged (provided they have *cos* sites).

4.4.2 Cloning in lambda

1. **Insertion vectors**. These are the simplest vectors. DNA is inserted into a single restriction site, which must be in a non-essential gene to maintain phage viability. Because of the upper limit on the size of DNA that can be packaged in the lambda head, there is a corresponding limit of a few kilobases on the amount of extra DNA that can be included. Lambda gt10 and gt11 are examples of insertion vectors and illustrate the limited amount of DNA they can accept (Figure 4.9). Lambda gt10 has a unique *Eco*RI site, within the *cI* repressor gene, and can accept inserts up to 7.6 kbp in size. This is a bit larger than inserts that wild-type phage would be able to accept, because there is a small deletion elsewhere in the genome of gt10. The presence of inserts can be detected by the fact that they inactivate the *cI* gene (see below). Lambda gt11 contains a *lacZ* gene with a unique *Eco*RI site. Protein-coding DNA sequences inserted into this can be expressed as a fusion with the LacZ protein (see Chapter 6). (The *cI*/*cro* control region of gt11 has, in fact, been replaced by that from another phage, 434, but the organization and expression of the gt11 genome both remain the same.)

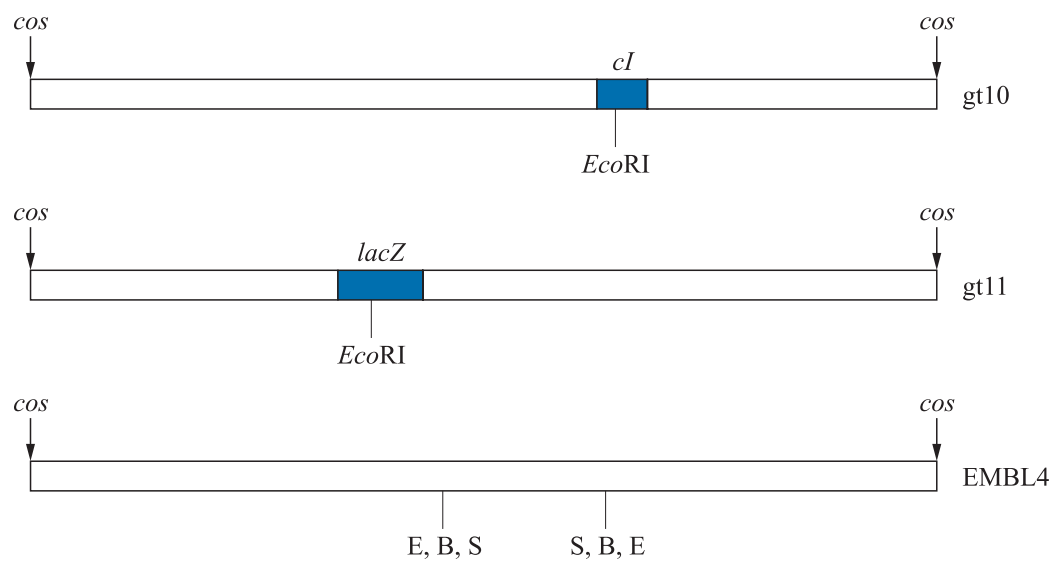


Fig 4.9 Example of lambda phage vectors. In gt10 and gt11 there are *Eco*RI sites in the *cI* and *lacZ* genes respectively (shaded) for insertion of DNA. EMBL4 is a substitution or replacement vector, from which a substantial non-essential region of the molecule can be removed by digestion with *Eco*RI (E), *Bam*HI (B) or *Sal*I (S).

2. **Substitution or replacement vectors.** In these vectors, a portion of the phage genome is removed by restriction digestion and is replaced with the DNA to be cloned. Much of the central region of lambda is not necessary for lytic growth and can be removed. The region includes the repressor gene, so the lysogenic mode of growth is no longer possible. The regions left include those responsible for phage DNA replication, the phage coat and its assembly, and cell lysis. About 20 kbp of the phage genome can, therefore, be removed and replaced with a piece of inserted DNA, which is called the stuffer fragment. The stuffer fragment is often generated by partial digestion with an enzyme with a 4 bp recognition site (see Chapter 1). Figure 4.9 shows a typical substitution vector, EMBL4. It contains two copies of a multiple cloning site, flanking the region to be removed. Digestion with an enzyme cutting within the multiple cloning site will separate the central region and the left and right flanking sections or arms.

   Before we can ligate in the stuffer fragment, it is necessary to take steps to stop the central portion from being religated to the arms again. This can be done by cleavage of the central portion with a second enzyme or by physical separation. If a second enzyme is used, it must cut within the central region, but not in the arms, so *Sal*I could be used for EMBL4 previously cut with *Bam*HI (Figure 4.9). It may not be necessary to remove the fragments generated, because the multimolecular reaction needed to regenerate the original phage would be so infrequent. Alternatively, if the second cuts are at sites very close to the first, such as with a *Bam*HI digest of EMBL4 followed by a *Sal*I digest, then the small fragments generated by the second digestion may be too small to be precipitated by ethanol. Therefore, precipitation of the DNA after the second cut and recovery of the pellet would yield the arms (with *Bam*HI ends), along with a part of the central region (with *Sal*I ends) that could no longer be ligated into them.

   Physical separation can be done either in an agarose gel or by centrifugation in a sucrose density gradient (which separates DNA molecules by size, in contrast to centrifugation in a caesium chloride gradient, which separates DNA molecules by their density). The arms are then recovered from the gel or the gradient. Generally, the DNA recovered from the gel is less efficient in ligation reactions than that recovered from sucrose gradients, so the latter technique is preferred.
3. **Packaging in vitro.** It is possible to introduce recombinant lambda molecules into a host by transformation with naked DNA, as we do with plasmids. However, this process is very inefficient, and we can improve it by orders of magnitude if we package the DNA in vitro into phage coats and use the phage's normal DNA infection process to get the DNA into the host cell. Packaging in vitro is essentially just a matter of incubating lambda DNA in a concatemeric configuration with a lysate of lambda-infected cells.

This lysate contains the lambda proteins (and host proteins such as chaperones) needed for phage assembly and should, therefore, be able to package the added recombinant DNA too. However, in practice it is not that simple, as there would be a high background of non-recombinant lambda phage produced from the packaging extract (which would contain lambda DNA as well as the packaging proteins). There are two approaches to making packaging extracts. One is to use two separate strains, carrying lambda prophage with chain-termination mutations in different genes for coat components (often the *D* and *E* genes). It is convenient if the prophages have a temperature-sensitive CI repressor, so that they can be induced by heat shock. On induction, packaging proteins are produced in each strain; however, no packaging can occur, as neither strain produces the full complement of proteins needed. Lysates of the cells from both strains are prepared and mixed. The mixed lysate has all the strains needed for packaging, which can now start. At this point, the DNA we want to package is added.

It is important to use suitable strains for making packaging extracts. They generally carry prophage with the following features:

(a) **Amber chain-termination mutations in coat protein genes**, usually mutations in the *D* and *E* genes.

(b) **The *cI*^{857} temperature-sensitive repressor protein gene**, which allows control of prophage induction by heating.

(c) **Chain-termination mutations in the *S* gene**. These mutations block cell lysis and allow the growth of cells containing very high levels of packaging proteins. We can then lyse the cells artificially.

(d) ***Red* mutation**. This mutation blocks a phage-encoded recombination pathway and, therefore, decreases the likelihood that recombination (and rearrangement) of DNA will take place in vitro during packaging. The lysogenic strains also usually carry the *recA* mutation, which largely blocks the *E. coli* recombination pathway and further reduces the likelihood of rearrangement in vitro.

(e) **A deletion in the *b* region**. This prevents excision of the prophage on induction and, therefore, reduces the amount of endogenous phage DNA in the packaging mix and the consequent level of background non-recombinant phage after packaging.

The second approach to making packaging systems uses a single lysogenic strain that will produce all the proteins needed for coat protein assembly. Using a single strain might be expected to lead to an increased likelihood of endogenous DNA being packaged. This is counteracted by having (as well as the mutations listed above) additional mutations, such as *xis*, in the prophage to stop the excision and packaging of the lysogenic phage from the cells producing the packaging extract.

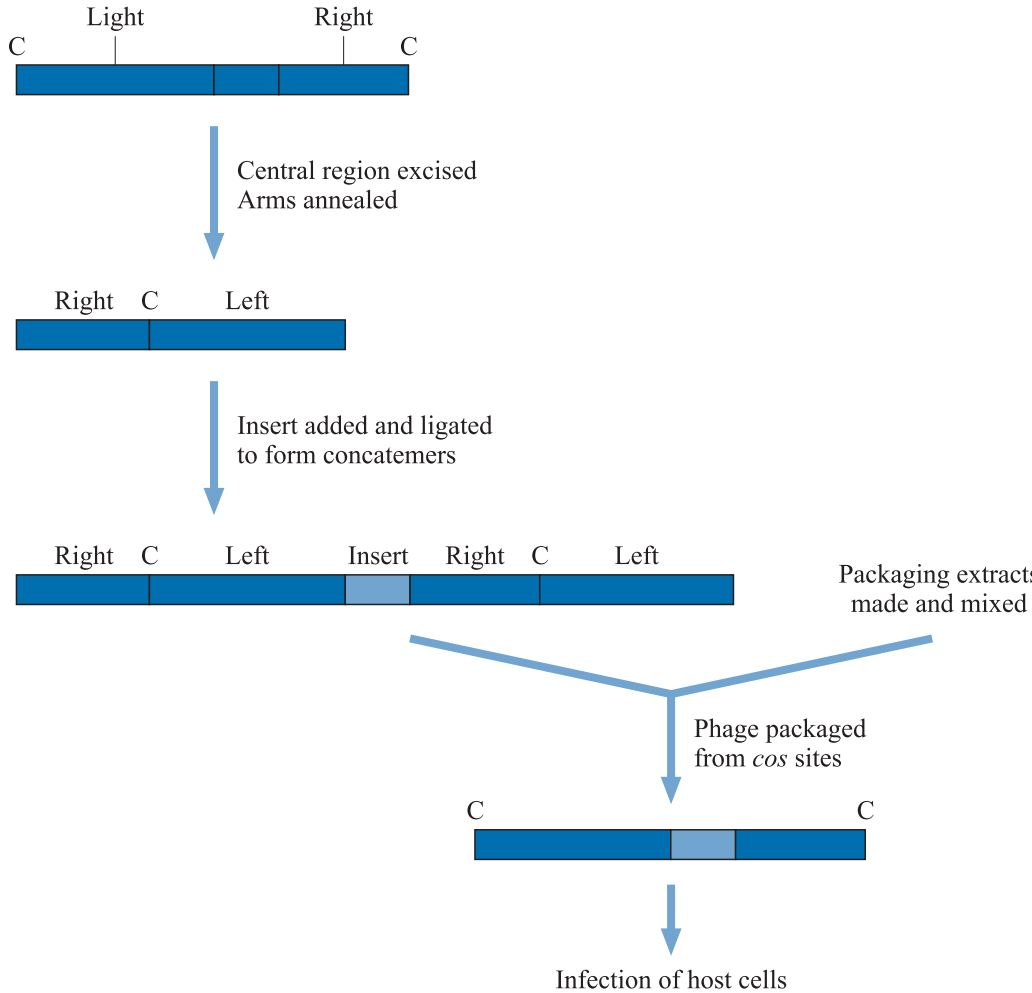


Fig 4.10 Summary of cloning into a lambda substitution vector. C = *cos* site.

Because packaging phage molecules requires the presence of *cos* sites on each side of the region to be packaged, the optimal substrate for packaging is a (left arm–insert–right arm)$_n$ concatemer, and the molar ratio of arms to insert should, therefore, be adjusted appropriately for the ligation. This is usually done empirically by carrying out small-scale ligation and packaging reactions. When the optimum ratio is established, a large-scale reaction is carried out. Note that the left and right arms are held together in the concatemer by annealing at the *cos* site. Once packaged, the phage DNA is introduced into *E. coli* by infection, as with lambda phage generated by more conventional means. The whole process is summarized in Figure 4.10.

4. **Identification of recombinants.** As with M13 cloning, uptake of a phage DNA molecule is marked by the formation of a plaque on an *E. coli* lawn. It is necessary to be able to distinguish between phages carrying an insert and those lacking one (where

the central portion has not been removed). There are many ways of going about this, depending on the particular phage used. They include:

(a) **Screening for *cI* function**. The CI repressor protein is required for the formation of lysogens. Phage in which the *cI* gene has been inactivated, therefore, are unable to form lysogens and will produce clear plaques rather than turbid ones. (Plaque turbidity is caused by the presence in the plaque of lysogenic cells that are immune from further infection because of the CI protein in their cytoplasm.) So, if the cloning site (in an insertion vector) or the removed region (in a replacement vector) involves the *cI* gene, then recombinant phage will be cI^- and can be recognized by their forming clear plaques, whereas non-recombinant phage will be cI^+ and form turbid plaques. This screening can be enhanced using an *hfl*A (high frequency of lysogenization) mutant host, in which CII stability is enhanced. This causes lysogenization to take place so efficiently that phage capable of forming lysogens do so in preference to entering the lytic pathway and fail to form plaques at all. Only phage that are incapable of forming lysogens (i.e. those that have lost *cI* function) will form plaques.

(b) ***spi* selection**. The use of *spi* selection relies on a slightly complex piece of phage biology. The *red* and *gam* gene products of phage lambda (an exonuclease involved in recombination and a protein inactivating the *E. coli* RecBCD nuclease respectively) inhibit its growth in *E. coli* cells that are lysogenic for the bacteriophage P2. However, red^-gam^- phage can grow in a P2 lysogen. Because the *red* and *gam* genes are removed during cloning in replacement phages, phage that have acquired an insert will be able to grow on a lawn of P2-lysogenic *E. coli*, and those that have not acquired an insert will retain *red* and *gam* and be unable to grow on the P2 lysogen. Recombinant phage can, therefore, be selected by plating on a P2 lysogen. (For propagation of the phage for subsequent DNA preparation, a host lacking the P2 prophage is used, as growth is much better.) Phage that are red^-gam^- are often designated spi^-, and others denoted spi^+ (sensitive to P2 inhibition).

Unfortunately, red^-gam^- phage produce rather few virions in infected cells. This is because the RecBCD nuclease (which is inhibited by the Gam protein) blocks rolling circle replication. Hence, gam^- phage are unable to inhibit the action of RecBCD and, therefore, are unable to carry out rolling circle replication. Consequently, the only way concatemers (which are needed for the production of packaged phage) can be formed is by recombination between the circular molecules formed by theta replication. Recombination between two circular molecules would generate a circular dimer, which is sufficient for packaging. However, the host recombination enzymes may not

work on lambda DNA (if it does not contain the necessary recombination initiation sites, called *chi* sites) and the phage recombination system (which relies on the *red* gene) is also missing. The only way a useful level of recombination (and concatemer formation) can be ensured is to include the *red* gene or a *chi* site in the phage. Many phage, such as EMBL4, have *chi* sites. The insert DNA may by chance also have them, but it is unwise to rely on this; phages with inserts that lack a *chi* site will be at a selective disadvantage.

(c) ***Lac* screening**. Some insertional vectors have a cloning site in a *lacZ* gene that has been introduced into the phage genome. Insertion of DNA into the cloning site, therefore, inactivates the production of beta-galactosidase in infected cells, and plaques on a lawn on medium containing X-gal and IPTG will be colourless. Plaques from phage lacking an insert will be blue. A similar approach can be used with replacement vectors, if the region replaced carries the appropriate part of a *lacZ* gene. Phage with an insert will give colourless plaques; those without will give blue plaques. The *lacZ* can be either a minigene (as with the pUC vectors described in Chapter 3) or an intact gene.

5. **Amplification**. Amplification is often useful when constructing libraries in lambda, especially if the number of phage produced after packaging in vitro is low. The procedure, which should not be confused with chloramphenicol amplification of plasmids (Chapter 3), is simply the propagation of the phage in *E. coli*. Each phage that infects an *E. coli* cell will give rise to many more phage after the cell is lysed, so the result is an increase in the number of phage present. This procedure can be carried out several times in succession, to bulk up the number of phage enormously. Note, though, that it does not increase the number of *different* sequences that have been cloned. Sometimes it may do the reverse. Phage carrying certain inserts may replicate more slowly than others. They will, therefore, be replicated fewer times during amplification and will, as a result, become underrepresented. In extreme cases, they may be lost completely, so the more times a library is amplified, the less representative it may become.

4.4.3 In vivo cloning in lambda

It is important to point out that lambda (and some other phages) can be used for cloning purposes without the need for restriction enzymes, ligase and so on. Induction of a lambda lysogen sometimes leads to aberrant excision of the phage from the *E. coli* chromosome. A molecule containing part of the phage DNA and part of the adjacent sequence is excised, rather than a single phage genome. Although the excised molecule itself may not contain all the phage genes for head proteins, tail proteins and so on, these genes are still present in the cell and the proteins will be produced. This results

in packaging of the phage, which now carry *E. coli* DNA replacing some of the phage DNA. Because of the limits on the size of molecules that can be packaged, phage particles that have acquired significant amounts of flanking sequence must have lost some phage sequence. These phage particles are called transducing phage; under the appropriate conditions, they can be selected and harvested, and their DNA can be obtained. Lambda is a specialized transducing phage, as it has a preferred site of integration into the *E. coli* chromosome (the attachment site) and will, therefore, preferentially produce transducing phage for genes near that site. If the attachment site is deleted, though, the phage can integrate elsewhere at secondary attachment sites, and transduce other genes. Any phage that can acquire chromosomal DNA (not necessarily by direct integration) can be used in transduction. Phages that do not preferentially transduce particular regions of the genome are called generalized transducing phages. The F factor can also be used, since it can integrate and excise, at times aberrantly, from the genome, to produce an F′. This process is often called sexduction. Both transduction and sexduction have proved enormously useful in obtaining *E. coli* genes. For more information, consult textbooks of classical bacterial genetics.

4.4.4 Lambda ZAP

The lambda ZAP vectors are based on bacteriophage lambda, but contain a region that can be excised in vivo to form one of the Bluescript plasmids (see Section 4.3.4). This is shown in Figure 4.11. The Bluescript region, which also contains the cloning sites, is flanked by two filamentous phage replication signals, the f1 initiator and the f1 terminator. The initiator is recognized by the gene II protein that is produced from a suitable helper phage. The site is nicked, and synthesis of a single strand is initiated. Replication proceeds unidirectionally from this site through the Bluescript region until the terminator is reached, at which point nicking takes place again. The single-stranded sequence that is generated is then circularized in vivo to form a covalently closed single-stranded circular molecule, which can then be converted into a double-stranded molecule by cellular DNA synthesis.

The double-stranded molecule that has been generated is the Bluescript plasmid and can be replicated as a plasmid from its ColEI origin of replication. Alternatively, in the continued presence of

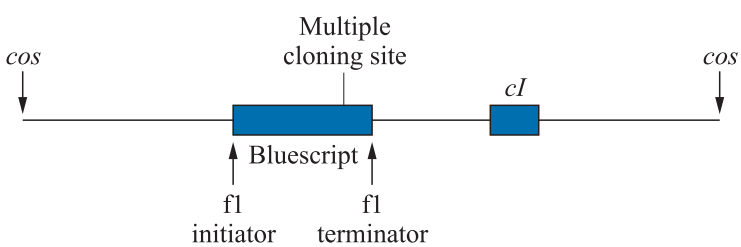


Fig 4.11 Lambda ZAP.

helper phage it can be used to synthesize single-stranded DNA, since it also contains an f1 origin of replication (which is a hybrid between the two origin sequences in the original lambda ZAP). The single-stranded DNA can also be packaged using the coat proteins produced by helper phage. As with Bluescript, the presence of an insert is detected by inactivation of *lacZ′*. So lambda ZAP can be used to construct a phage library, with inserts of up to 10 kbp. Once a suitable phage has been selected from the library, that phage is used to infect *E. coli* cells that are co-infected with the f1 helper phage. This results in Bluescript plasmid excision. After a few hours' growth, the excised and packaged DNA is recovered. It is then used to reinfect *E. coli*. On entering *E. coli*, it is converted to dsDNA by the cell, giving rise to ampicillin-resistant *lacZ*$^-$ colonies, and is maintained as a double-stranded plasmid. This procedure has, in effect, allowed subcloning from the lambda phage directly into a plasmid in vivo. Otherwise, this would have to be done much more laboriously by excision of the insert from the phage using restriction endonucleases, followed by ligation into a suitably cut vector.

4.5 Cosmids

Cosmids can be regarded as lambda replacement vectors from which even more phage DNA has been deleted, to leave only the *cos* packaging site. They also contain a plasmid origin of replication. None of the coat protein genes is left; so, although cosmids can be packaged in vitro, once they are inside an *E. coli* cell they cannot generate more phage and can, therefore, propagate only as plasmids. They cannot give rise to plaques. For this reason, some other selectable marker must be used to indicate their presence, such as a beta-lactamase gene for ampicillin resistance (Figure 4.12). Entry of DNA into the host cell is selected not by the formation of a plaque on a lawn, but by the formation of colonies on plates containing ampicillin. Cosmids are, in effect, bacterial plasmids that happen to be able to be packaged into phage coats for easy delivery.

The procedure for inserting DNA into cosmids is very similar to the one used with lambda. Cosmids are much smaller than lambda phage. However, the same minimum-size condition for packaging and delivery to *E. coli* pertains. Hence, the only cosmids that can be successfully packaged, delivered and form ampicillin-resistant colonies are those that have acquired large inserts. There is, therefore, a selection for the acquisition of large inserts, and this is the main advantage of cosmids. Typically, they will accept inserts of 35 to 45 kbp. They are, therefore, popular for the construction of genomic libraries (see Chapter 5). However, a drawback is that once inside the *E. coli* cell there is no longer any size selection, and partial deletion may take place. The tendency for deletion can be avoided using

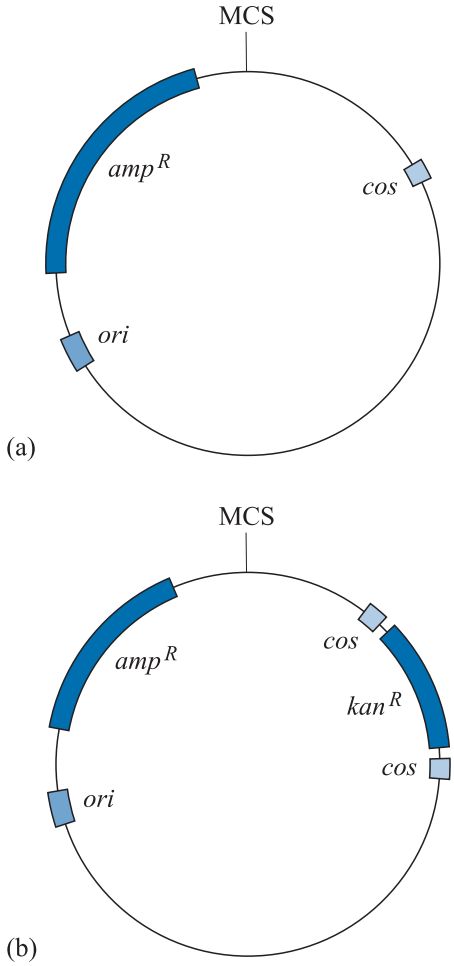


Fig 4.12 Cosmids. (a) Cosmid pJB8 (5.4 kb) contains an ampicillin resistance gene (*amp*^{R}), an *E. coli* origin of replication (*ori*), a multiple cloning site (MCS) and a *cos* site. (b) Cosmid c2RB (6.8 kb) contains a second *cos* site (and an additional selectable marker, *kan*^{R}, for kanamycin resistance).

vectors with lower copy number. This has led to the generation of cosmids where the ColEI-derived plasmid origin of replication is replaced by an origin of replication from the F plasmid (which is present at a lower copy number). These vectors are sometimes referred to as fosmids. Some cosmids contain two *cos* sites, e.g. c2RB; see Figure 4.12. This makes it unnecessary to ensure concatemer formation before packaging, as each molecule already has a pair of *cos* sites.

4.6 Bacteriophage Mu

Bacteriophage Mu is another example of a temperate phage. It is packaged into heads containing 39 kbp of DNA. When the phage replicates, it undergoes transposition hundreds of times, inserting its genome at multiple sites in the host genome. Mu can be used for cloning in vivo, as a 'mini-Mu' that contains a reduced phage genome to which have been added selectable markers and a plasmid

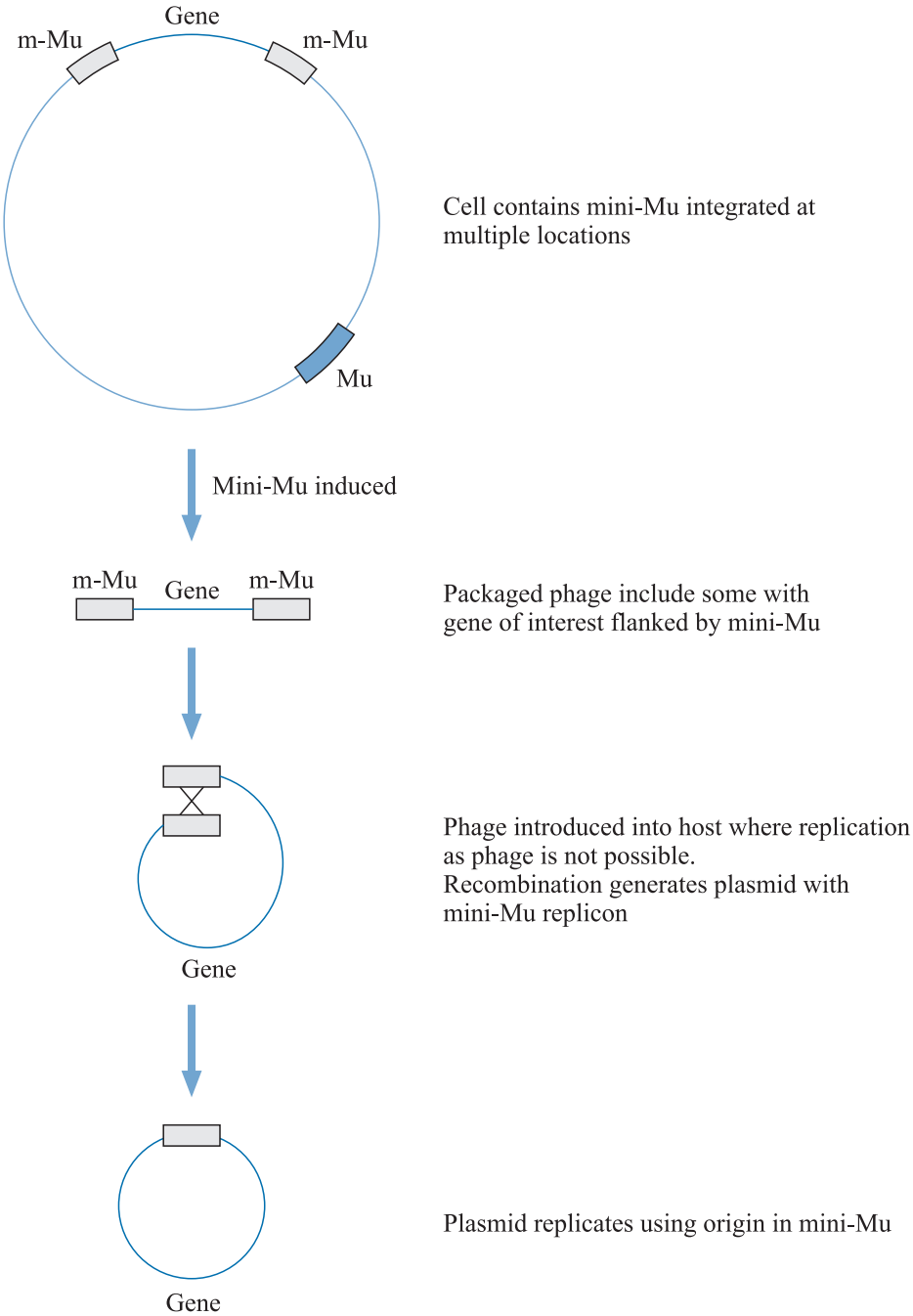


Fig 4.13 Cloning in vivo using mini-Mu. The grey box represents the mini-Mu element (m-Mu), which contains an origin for plasmid replication and a selectable marker. The initial host strain contains an additional Mu prophage (blue box) to provide packaging proteins.

replication origin. The basis of the method is shown in Figure 4.13. A culture of the target strain, whose DNA we want to clone, is infected with mini-Mu. The phage inserts at multiple sites throughout the genome, prior to excision and packaging. (Packaging functions are provided by a conventional Mu prophage that is also integrated into the target strain's genome.) Among the packaged phage that are produced from the infected culture will be some that contain two mini-Mu genomes flanking a region of DNA from the target strain. The packaged phage are introduced into a second, recipient, strain in which Mu is unable to replicate as a phage. In cells of the second strain that receive a molecule containing two mini-Mu elements, recombination between the elements can occur, yielding a circular molecule that can replicate as a plasmid from the origin carried by the mini-Mu element. Therefore, we have a plasmid that contains a piece of DNA from the first target strain and which is stably maintained in the second strain. Although this may seem to rely on a number of rather improbable events (insertion of two mini-Mu elements flanking the gene of interest in the target strain, recombination between the two elements in the second host), because of the ease of handling very large numbers of phage particles, it can be a very effective method of cloning. Clearly, it is restricted to cloning fragments of genomic DNA from bacterial species or strains in which phage Mu can replicate.

4.7 Bacteriophage PI

Although the use of cosmids allows large molecules, up to 40 kbp or so, to be cloned, it is often desirable to be able to clone even larger fragments. We have seen that BACs can be used for this. Bacteriophage P1 may also be useful for cloning larger fragments, as its head can accommodate some 110–115 kbp. The vectors used are complex, as a

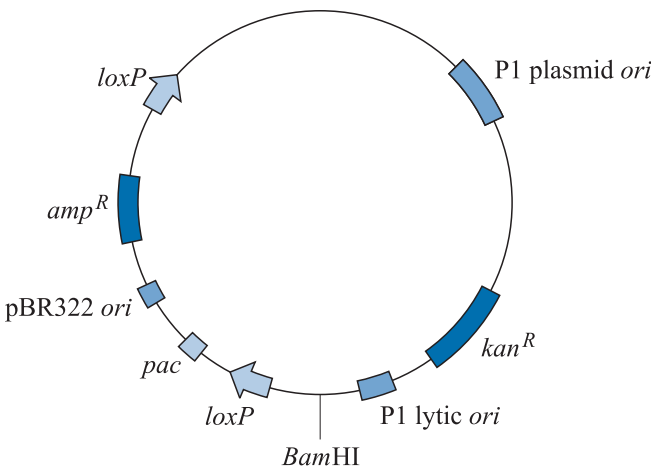


Fig 4.14 The bacteriophage PI cloning vector pNS582 (16.0 kb). The features are described in the text.

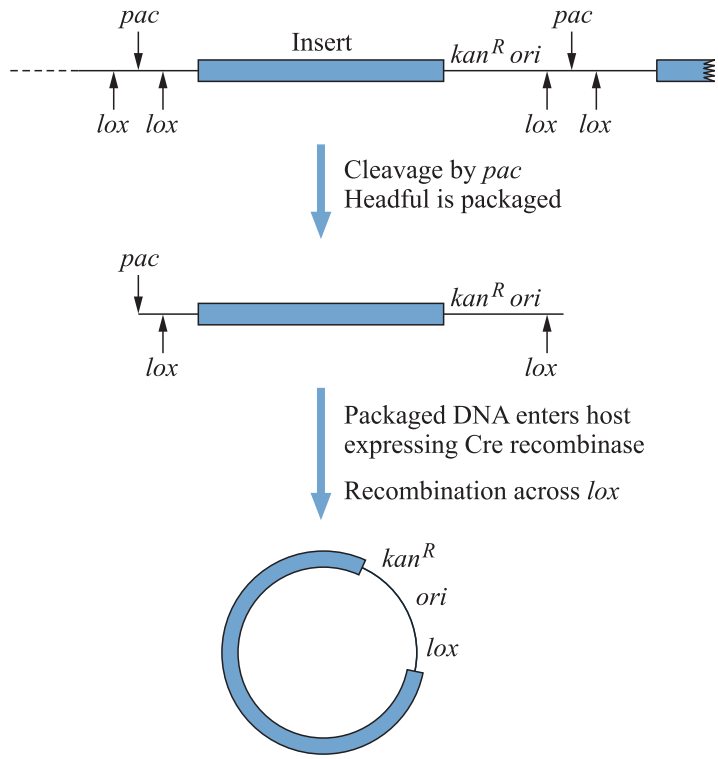


Fig 4.15 Cloning using a PI vector. The method and the features of the vector are described in the text. Only the PI plasmid origin is shown. The pBR322 and PI lytic origins, and amp^R have been omitted for simplicity.

consequence of the requirements for packaging, and an example is given in Figure 4.14. The *pac* site is involved in the initiation of packaging of DNA into phage heads, and the *loxP* sites in the cyclization of linear molecules in a host bacterium. The cloning procedure is as follows. The DNA is prepared by growth of the vector as a plasmid in *E. coli* using the pBR322 origin, and then cleaved with *Bam*HI. Then *Sau*3A *partial* digestion products are ligated in, which will produce molecules such as those in Figure 4.15 (among others). The next stage is to incubate the molecules with an extract of *E. coli* obtained by induction of a suitable strain lysogenic for a mutant P1. This mutant lysogen produces proteins that cleave the recombinant molecules at the packaging site *pac*, but it does not produce the heads and tails for packaging. The DNA is then incubated with a different extract that this time contains head and tail subunits but not the proteins that cleave at *pac*. This brings about packaging of DNA into heads, starting at the *pac* end and then progressing until a full head has been packaged. At this point, a sequence-independent cleavage of the DNA takes place. (Molecules with too short an insert will not contain sufficient DNA to constitute a phage headful before the next *pac* site is reached.) The packaged DNA can now be introduced into bacterial cells by infection, but circularization of the DNA must take place inside these cells before it can be replicated. In vectors based on bacteriophage lambda, this would be brought about simply by annealing the *cos* ends of the molecule, but this is not possible in the P1 system, as the sequences at the ends will not be complementary.

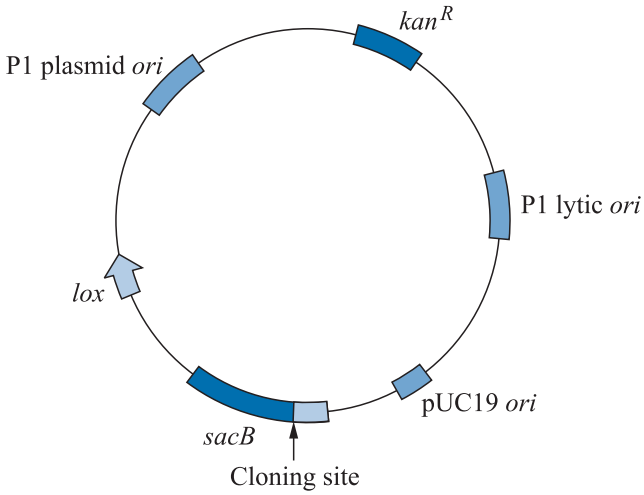


Fig 4.16 The PAC vector pCYPAC-1 (20.0 kb). The *sacB* gene encodes levansucrase, allowing selection for the presence of inserts.

(Remember that cleavage after a P1 headful has been packaged is sequence independent.) Instead, circularization takes place by recombination between the *loxP* sites, which is mediated by the *cre* gene product (usually a P1 phage gene product, but which is constitutively expressed in the host strain used for this cloning). The circular molecule produced can then replicate stably as a plasmid. There are in fact two P1 replication origins on the vector. One is for routine maintenance as a low-copy plasmid, and the other (the 'lytic' origin) can be artificially activated for replication at high copy number prior to recovering the recombinant DNA from cells carrying it.

The P1 vectors have been further developed by removal of the phage packaging signals to produce vectors that still have the two P1 replication origins, but which are introduced into the host cell by electroporation rather than packaging (Figure 4.16). Therefore, they rather resemble the BAC vectors we discussed at the start of the chapter and are called PAC vectors. In the absence of packaging constraints they can accept inserts of up to 300 kbp. Maintenance at low copy number helps to avoid instability, and the copy number can be increased by activation of the appropriate origin. The cloning site lies between an *E. coli* promoter and a promoterless copy of the *sacB* gene from the bacterium *Bacillus amyloliquefaciens*. This gene encodes levansucrase, which hydrolyses sucrose to glucose and fructose, and polymerizes the latter to form the compound levan, which is toxic to *E. coli*. Thus, expression of *sacB* is lethal in the presence of sucrose in the growth medium. The presence of an insert in the PAC cloning site between the promoter and the *sacB* gene abolishes transcription of the gene and production of sacB, allowing cells to survive in the presence of sucrose. If the insert is lost, or not present initially, cells will be killed. Direct selection for the presence of an insert in this way is called positive selection. Selection for the *absence* of a marker (*sacB* in this case) is called counter-selection.

Chapter 5

Making libraries

5.1 Introduction

We encountered the concept of a library, a collection of random DNA clones, in Chapter 3. Having learned about cloning in bacteriophage viruses and their derivatives in Chapter 4, we are now in a position to look in more detail at making libraries.

Libraries can conveniently be divided into two categories: genomic libraries, which are made from the total genomic DNA of an organism, and cDNA libraries, which are made from DNA copies of its RNA sequences. We will look first at how these two types of library are made. Sometimes, though, it is useful to be able to make a more specialized genomic or cDNA library, enriched for particular sequences, so we will look at how they are made too. The library we looked at in Chapter 3 was a plasmid genomic library. It was represented by a large number of colonies on a plate, each containing a plasmid with a defined insert. We can also use a phage vector; handling large numbers of phages is often more convenient than handling large numbers of plasmids. When using phage, we get a collection of plaques on a lawn of bacterial cells, rather than colonies. Each plaque will contain a single type of phage with a defined DNA insert. With cosmids, fosmids, BACs and PACs, replicating as plasmids, the library will again be colonies on a plate. With yeast artificial chromosomes (YACs) and mammalian artificial chromosomes (MACs), described in Chapter 9, the library will be propagated in yeast or mammalian cells. Many companies sell 'ready-made' genomic and cDNA libraries. Although these can be expensive, it is often more cost effective to use such a library rather than spend a lot of time preparing your own.

5.2 Genomic libraries

5.2.1 Principles

A genomic library contains all the sequences present in the genome of an organism (apart from any sequences, such as telomeres that

cannot be readily cloned). Clearly, the larger the insert of genomic DNA in each recombinant, the lower the number of recombinants needed to represent the organism's genome completely. (The relationship between genome size, insert size and the number of members needed for the library is discussed in more detail in Chapter 6.) For most purposes, therefore, it is best to use vectors that will accept large inserts. This effectively means lambda replacement vectors, cosmid and fosmid vectors, BACs, P1 and BAC vectors. YACs, which are discussed in Chapter 9, are also widely used, as they can also accept very large inserts. For small genomes, plasmids or lambda insertional vectors may be suitable.

5.2.2 The procedure

1. **Preparing DNA.** The key to generating a high-quality library usually lies in the preparation of the insert DNA. The first step is the isolation of genomic DNA. The procedures vary widely according to the organism under study, and will not be discussed here. Care should be taken to avoid physical damage to the DNA so that it is of as high a molecular weight and as free of nicks as possible. If the intention is to prepare a nuclear DNA library, then it is often sufficient to use total DNA, ignoring whatever DNA is present in the mitochondria or chloroplasts, as there is usually much more nuclear than organellar material (in total amount, but not necessarily in numbers of copies of individual sequences). If the aim is to make an organelle genomic library, then it would be wise to purify the organelles away from the nuclei first and then prepare DNA from them.
2. **Fragmentation of DNA.** The DNA is then fragmented to a suitable size for ligation into the vector. As discussed in Chapter 1, this could be done by complete digestion with a restriction endonuclease, but a large number of sequences would not be represented intact (or might not be represented at all). It is much better to use partial digestion with a frequently cutting enzyme (such as *Sau*3A, with a four-base-pair recognition site) to generate a random collection of fragments with a suitable size distribution. If one is using a vector that takes larger inserts, then it may be easier to use partial digestion with an enzyme with a six-base-pair recognition site. Once prepared, the fragments that will form the inserts are often treated with phosphatase, to remove terminal phosphate groups. This ensures that separate pieces of insert DNA cannot be ligated together before they are ligated into the vector. Ligation of separate fragments is undesirable, as it would generate clones containing non-contiguous DNA, and we would have no way of knowing where the joint lay. Note that this is a different use of phosphatase from the one encountered earlier, where the *vector* (rather than the insert) was treated with phosphatase to stop it self-ligating.
3. **Vector preparation.** This will depend on the kind of vector used. The vector needs to be digested with an enzyme (or enzymes) appropriate to the insert material we are trying to clone.

4. **Ligation and introduction into the host.** Vector and insert are mixed, ligated, packaged (if appropriate) and introduced into the host by transformation, infection or some other technique.
5. **Amplification.** This is not always required. Libraries using phage cloning vectors are often kept as a stock of packaged phage. Samples of this can then be plated out on an appropriate host when needed. Libraries constructed in plasmid vectors are kept as collections of plasmid-containing cells, or as naked DNA that can be transformed into host cells when needed. With storage, naked DNA may be degraded. Larger molecules are more likely to be degraded than smaller ones, so larger recombinants will be selectively lost, and the average insert size will fall.

5.3 cDNA libraries

5.3.1 Principles

For cDNA libraries, we produce DNA copies of the RNA sequences (usually the mRNA) of an organism and clone them. Such libraries are particularly useful because they represent not just the collection of expressed sequences from that organism, but also those sequences after any post-transcriptional modification, such as the removal of introns. Comparison of cDNA sequences with genomic DNA sequences allows the determination of the positions of introns, polyadenylation sites and so forth. Partial sequencing of a large collection of cloned cDNAs and deposition in a database is often the first stage in a genome characterization project for an organism. These cDNAs are often referred to as expressed sequence tags (ESTs).

The cDNA molecules to be cloned are often no more than a few kilobase pairs long, so plasmid vectors may well be suitable. However, lambda vectors such as lambda gt11 and lambda-ZAP, are also quite widely used, and may have advantages for the subsequent screening process. For a long time, cDNA cloning was regarded as being particularly difficult, and it was notoriously hard to generate *full-length* cDNA (i.e. cDNA molecules containing a complete open reading frame, and thus more or less the entire length of the RNA species). Things were made much easier by the development of the RNaseH technique described below.

The approach adopted for cDNA cloning depends in the initial stages on whether or not the RNA species for which cDNA is required are polyadenylated at their 3′ ends. Most eukaryotic cytoplasmic mRNAs are polyadenylated, by the addition after synthesis of a run of A residues. Those mRNAs that are not generally polyadenylated include those for histones, some organelle mRNAs (e.g. those from chloroplasts and plant mitochondria), bacterial mRNAs and stable mRNAs such as ribosomal RNAs.

5.3.2 Polyadenylated RNA

Polyadenylated RNA (polyA$^+$ RNA) can be separated from other RNAs by exploiting its ability to bind to oligo-dT (short oligonucleotides composed entirely of deoxyT residues). This is commonly done with oligo-dT cellulose; that is, cellulose (acting as a solid-phase support) to which oligo-dT has been covalently attached via the −OH goups of the cellulose. Usually, a solution containing the RNA is passed through a column of oligo-dT cellulose. The polyA tail of the RNA forms hydrogen bonds with the oligo-dT, and polyA$^+$ RNA, therefore, is retained by the column. After washing all non-specifically bound RNA from the column, the polyA$^+$ RNA is eluted with a low-salt buffer. (High salt concentrations stabilize nucleic acid hybridization, and low salt concentrations weaken it.) There are variations on this method. For example, the oligo-dT can be attached to streptavidin rather than cellulose. The oligo-dT–streptavidin conjugate (with polyA$^+$ RNA base-paired to the oligo-dT) can then be recovered using magnetic beads with biotin attached. The biotin binds to the streptavidin and the whole ensemble is recovered using the magnetic properties of the beads. Note that the principle of the method is still the interaction between the polyA tail of the message and oligo-dT.

Preparation of good quality polyA$^+$ RNA is probably the most important part of cDNA cloning, and it is made especially difficult by the fact that RNA is a particularly labile molecule (much more so than DNA, because the 2′-hydroxyl group of the ribose ring increases its reactivity), necessitating careful precautions against degradation. These usually include baking all glassware and treatment with UV light and ribonuclease inhibitors. Not only is RNA particularly labile, many ribonucleases are very stable, and some can be boiled with little subsequent loss of activity.

5.3.3 cDNA synthesis

1. **The RNaseH method.** The principle of this method is that a complementary DNA strand is synthesized using reverse transcriptase (Chapter 1) to make an RNA:DNA duplex. The RNA strand it then nicked and replaced by DNA (Figure 5.1). The first step is to anneal a chemically synthesized oligo-dT primer to the 3′ polyA-tail of the RNA. The primer is typically 10–15 residues long, and it primes synthesis of the first DNA strand with reverse transcriptase and deoxyribonucleotides. This leaves an RNA:DNA duplex, and the next step is to replace the RNA strand with a DNA strand. The difficulty is finding a way to prime synthesis using the DNA strand as template. Annealing oligo-dA to the oligo-dT incorporated during synthesis of the first strand would be no use: the oligo-dT is at the 5′ end of the template molecule, but synthesis must start at the 3′ end. There is no way to predict the sequence at the 3′ end. However, we can use as a primer the RNA that is already base-paired to the DNA. In practice, this is done by treating the

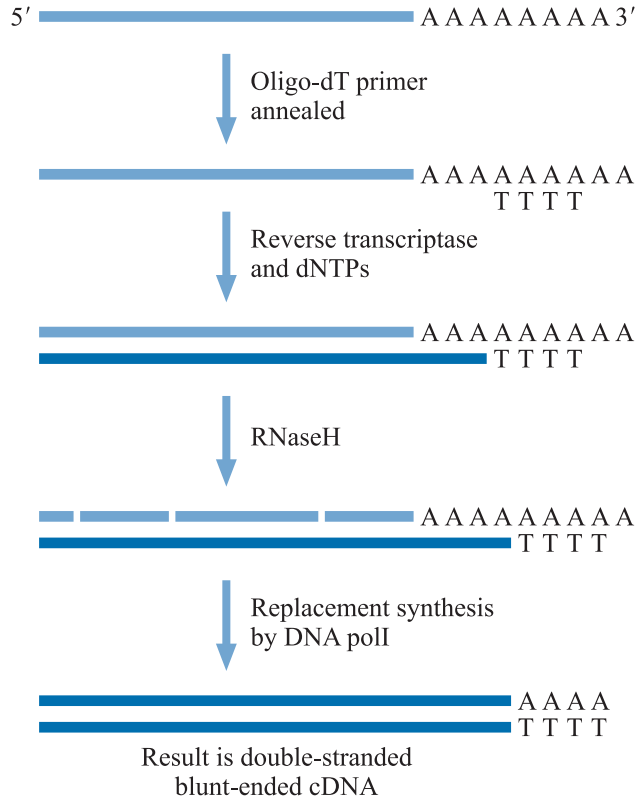


Fig 5.1 Synthesis of cDNA by the RNaseH method. Dark blue shaded material is DNA; light blue shaded material is RNA.

RNA:DNA duplex with a low concentration of RNaseH, together with DNA polI and deoxyribonucleotides. The RNase nicks the RNA, leaving free 3′-hydroxyl groups, and DNA can then be made using these as primers. (Remember that DNA synthesis in vivo uses RNA primers with free 3′-hydroxyls.) As DNA chains are synthesized, any molecules that are base-paired to the template further down are displaced or degraded by the polymerase using its 5′-3′ exonuclease activity. Eventually this leaves a DNA:DNA duplex with perhaps a small region of RNA including any 5′ cap at one end. This is removed by addition of RNase, and any residual overhangs on the DNA molecule are polished with DNA polymerase (Chapter 1) to leave blunt-ended dsDNA.

2. **Integrity of the 5′ end.** Before the development of the RNaseH method described above, the synthesis of the second DNA strand was particularly problematic. The approach most widely used was to degrade completely the RNA strand of the RNA:DNA hybrid, to leave a single-stranded DNA molecule. Chance complementarity between the sequence at the 3′ end of this molecule and a sequence elsewhere in the same molecule would allow it to form a hairpin structure, as shown in Figure 5.2. The 3′ end of this hairpin could then be extended in a DNA synthesis reaction with

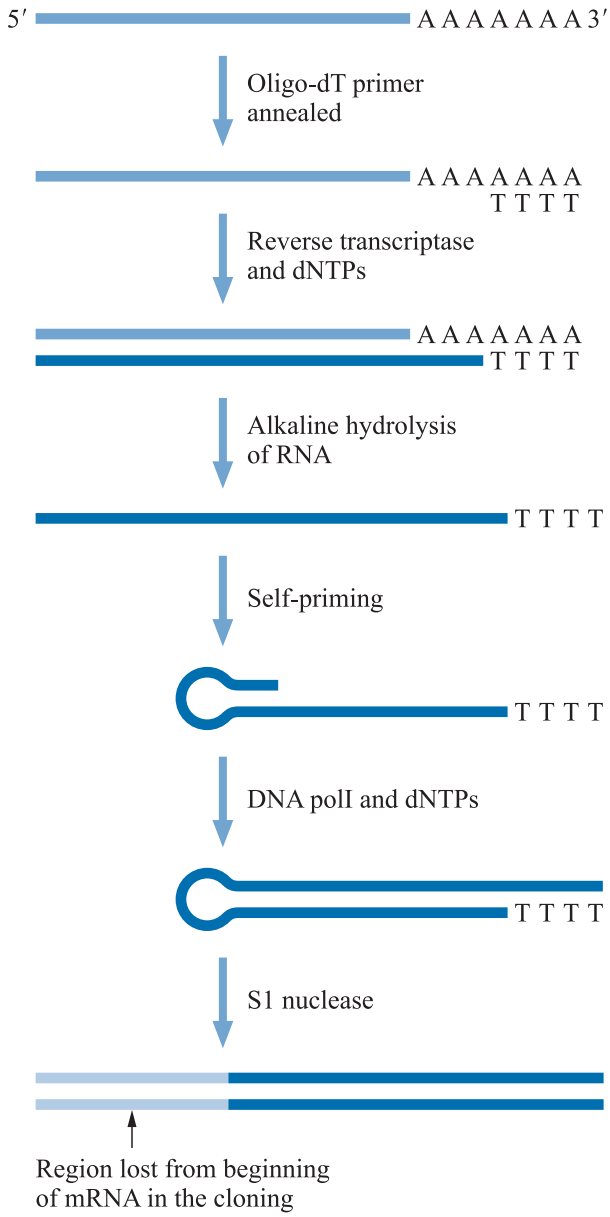


Fig 5.2 Synthesis of cDNA by the self-priming method. Dark blue shaded regions represent DNA; lighter shaded regions are RNA. The lightest shaded region represents material that is not represented in the final cDNA because of the loss of material in the self-priming.

DNA polymerase and deoxyribonucleotides. The loop region of the hairpin is then removed with a single-strand-specific DNase. There are two major problems with this approach. The first problem is that there is no guarantee that such a hairpin structure can form. The second is that much of the sequence corresponding to the 5′ end of the message will be lost. Thus, cDNA molecules synthesized by this means usually have large deletions extending in from the region corresponding to the 5′ end of the message. The RNaseH method largely avoided this problem and was a major improvement. However, even with this method, the cDNA may

lack part of the region corresponding to the 5′ end of the mRNA. Other methods have been derived to maintain the integrity of this region, which rely on homopolymer tailing.

3. **Homopolymer tailing.** This approach uses the enzyme terminal transferase (Chapter 1), which can polymerize nucleotides onto the 3′-hydroxyl of both DNA and RNA molecules. The method is illustrated in Figure 5.3. We carry out the synthesis of the first DNA strand essentially as before, to produce an RNA:DNA hybrid. We then use terminal transferase and a single deoxyribonucleotide to add tails of that nucleotide to the 3′ ends of both RNA and DNA strands. The result of this is that the DNA strand now has a known sequence at its 3′ end – an oligomer of whichever nucleotide was used in the tailing reaction. Typically, dCTP or dATP are used. A complementary oligomer (synthesized chemically) can now be annealed and used as a primer to direct second strand synthesis. This oligomer (and also the one used for first strand synthesis) may additionally incorporate a restriction site, to help in cloning the resulting double-stranded cDNA.
4. **RACE.** It is sometimes the case that we wish to clone a particular cDNA for which we already have some sequence data, but with particular emphasis on the integrity of the 5′ or 3′ ends. RACE techniques (Rapid Amplification of cDNA Ends) are available

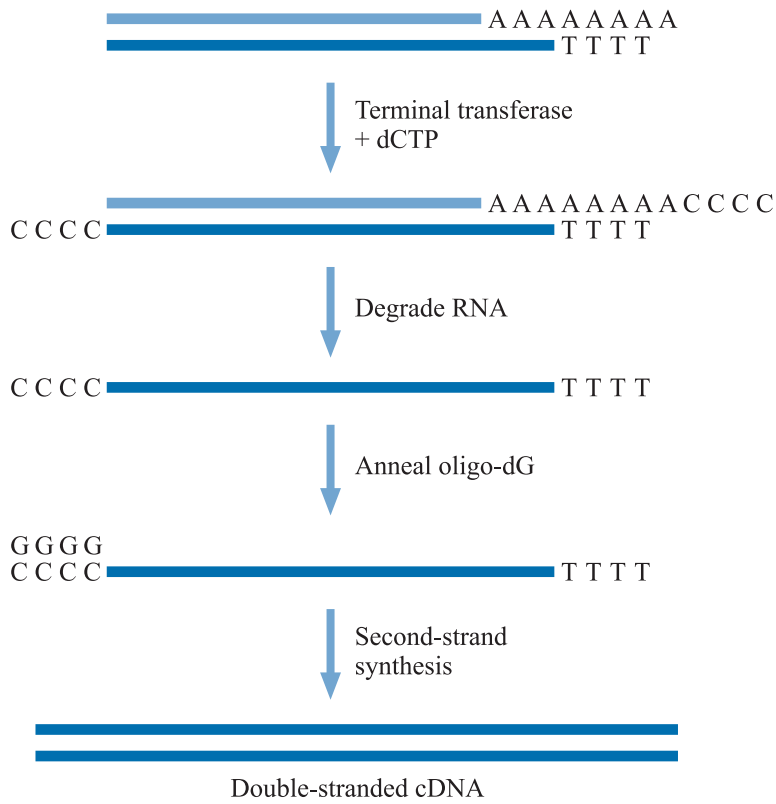


Fig 5.3 Homopolymer tailing to provide a priming site for second-strand synthesis. Dark blue shaded material is DNA; the light blue shaded material is RNA.

for this. The RACE methods are divided into 3′RACE and 5′RACE, according to which end of the cDNA we are interested in, and are shown in Figure 5.4. In 3′RACE, reverse transcriptase synthesis of a first DNA strand is carried out using a modified oligo-dT primer. This primer comprises a stretch of unique adaptor sequence followed by an oligo-dT stretch. The first strand synthesis is followed by a second strand synthesis using a primer internal to the coding sequence of interest. This is followed by PCR using (i) the same internal primer and (ii) the adaptor sequence (i.e. omitting the oligo-dT). Although in theory it should be possible to use a simple oligo-dT primer throughout instead of the adaptor-oligo-dT and adaptor combination, the low melting temperature for an oligo-dT primer may interfere with the subsequent rounds of PCR. In 5′RACE, a first cDNA strand is synthesized with reverse transcriptase and a primer from within the coding sequence. Unincorporated primer is removed and the cDNA strands are tailed with oligo-dA. A second cDNA strand is then synthesized with an adaptor-oligo-dT primer. The resulting double-stranded molecules are then subject to PCR using (i) a primer nested within the coding region and (ii) the adaptor sequence. A nested primer is used in the final PCR to improve specificity. The adaptor sequence is used in the PCR because of the low melting temperature of a simple oligo-dT primer, as in 3′RACE above. A number of kits for RACE are commercially available.

5.3.4 Cloning the cDNA

1. **Linkers.** The RNaseH and homopolymer tailing methods ultimately generate a collection of double-stranded, blunt-ended cDNA molecules. They must now be attached to the vector molecules. This could be done by blunt-ended ligation, or by the addition of linkers, digestion with the relevant enzyme and ligation into vector as described in Chapter 3. (Remember the need for methylase treatment if linkers are to be used.)
2. **Incorporation of restriction sites.** It is possible to adapt the homopolymer tailing method by using primers that are modified to incorporate restriction sites, as shown in Figure 5.5. In this example, the oligo-dT primer is modified to contain a restriction site (in the figure, a *Sal*I site GTCGAC). The 3′ end of the newly synthesized first cDNA strand is tailed with C's. An oligo-dG primer, again preceded by a *Sal*I site within a short double-stranded region of the oligonucleotide, is then used for second-strand synthesis. Note that this method requires the use of an oligonucleotide containing a double-stranded region. Such oligonucleotides are made by synthesizing the two strands separately and then allowing them to anneal to one another.

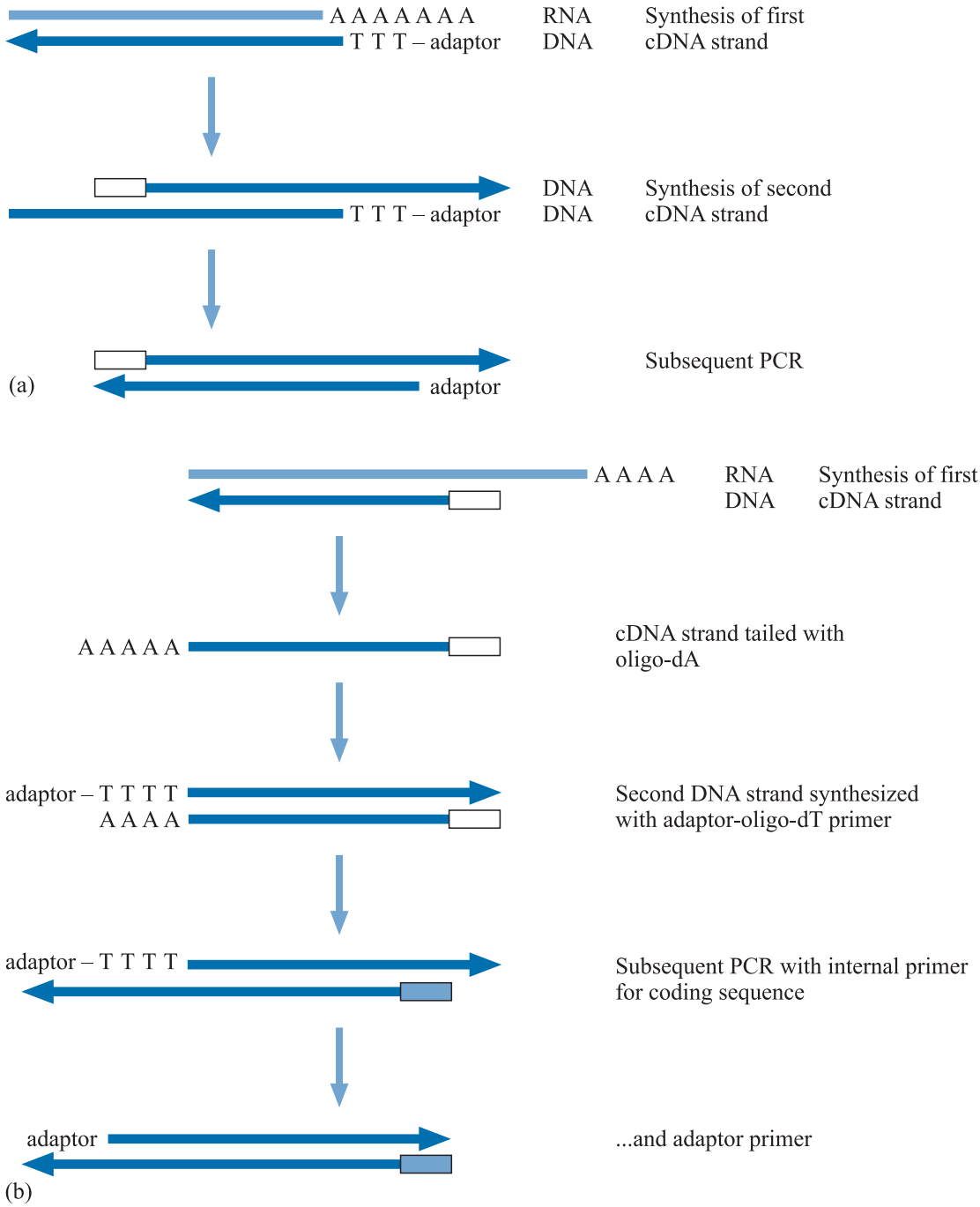


Fig 5.4 RACE. (a) 3′RACE. The first primer is the oligo-dT-adaptor molecule. The second primer (open box) is internal to the coding sequence of interest. This is used in conjunction with the adaptor primer (rather than the oligo-dT-adaptor primer) in subsequent PCR. (b) 5′RACE. Synthesis of the first cDNA strand uses a primer within the coding region (open box). The first cDNA strand is tailed with oligo-dA. A second DNA strand is synthesized with an adaptor-oligo-dT primer. This is followed by PCR with (i) a primer nested within the coding sequence (shaded box) and (ii) the adaptor.

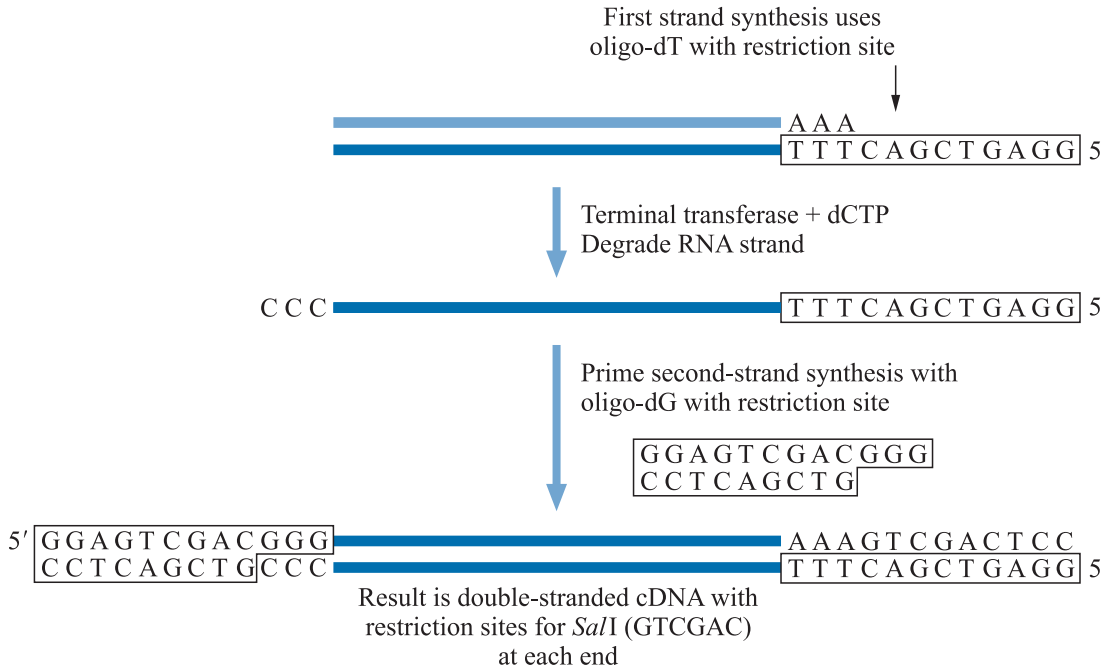


Fig 5.5 Modification of homopolymer tailing, incorporating restriction sites. The protocol is essentially the same as in Figure 5.3, except that the oligo-dT and oligo-dG primers also contain *Sal*I restriction sites. For clarity, the primers are shown boxed.

3. **Homopolymer tailing of cDNA.** Another option is to use terminal transferase again. Treatment of the blunt-ended double-stranded cDNA with terminal transferase and dCTP leads to the polymerization of several C residues (typically 20 or so) to the 3′ hydroxyl at each end. Treatment of the vector with terminal transferase and dGTP leads to the incorporation of several G residues onto the ends of the vector. (Alternatively, dATP and dTTP can be used.) The vector and cDNA can now anneal, and the base-paired region is often so extensive that treatment with DNA ligase is unnecessary. In fact, there may be gaps rather than nicks at the vector–insert boundaries, but these are repaired by physiological processes once the recombinant molecules have been introduced into a host.

5.3.5 Non-polyadenylated RNA

If the RNA is not polyadenylated, we cannot use oligo-dT to prime from the polyA tail. Instead, a collection of chemically synthesized oligonucleotides of random sequence is used. These are usually hexamers, and are produced by oligomerization of equal quantities of mixed A, G, C and T residues, so that all possible hexameric sequences should be present. They will, therefore, be able to bind throughout the length of the RNA molecule to prime synthesis of the first DNA strand. Synthesis of the second DNA strand is conveniently carried out with RNaseH and DNA polymerase I, as described above.

5.4 Specialized libraries

It is often helpful to make libraries that are enriched for a particular fraction of genomic DNA or cDNA. This may be the case if we are trying to clone a particular gene, for example, and have some limited information about, say, chromosomal location.

5.4.1 Shelves

Sometimes we know the size of the restriction fragment on which a particular gene is located. For example, this information may be acquired by probing a Southern blot of digested genomic DNA with a suitable sequence, such as an oligonucleotide probe, and measuring the size(s) of restriction fragment(s) that hybridize. Once the size of the relevant restriction fragment is known, another digest of genomic DNA is then carried out with the same enzyme. The products are separated by electrophoresis, and DNA fragments of approximately that size are recovered from the gel. They are then cloned into a suitable vector. Because they are likely to be smaller than the random fragments used in making full genomic libraries, a plasmid vector is often suitable. The collection of recombinants generated is frequently called a shelf, as it is a subsection of a library.

5.4.2 RNA selection

1. **Tissue type.** Any cDNA library is likely to represent only a fraction of the RNA species of any one organism, and this will depend on the particular type, developmental stage, physiological state and so on of the tissue from which the RNA was isolated. The library would, therefore, be expected to contain cDNAs for the general 'housekeeping' genes and for those genes whose expression is specific to that particular tissue. If we are interested in obtaining the cDNA for a specific protein, it would be wise to use a cDNA library from a tissue with a lot of RNA for that protein relative to other RNAs; in other words, we should use a cDNA library from tissue producing large amounts of the protein. By using micro-dissection or cell sorting techniques it may be possible to make libraries from just a few cells.
2. **RNA fractionation.** We have seen that fractionation on an oligo-dT-cellulose column can be used to separate polyA$^+$ RNA from polyA$^-$ RNA, such as rRNA, prior to the creation of the cDNA library. It may be useful to fractionate the mRNA further, to obtain a population enriched for the RNA for a particular protein. The procedure outlined here requires the availability of antibodies to the protein and is summarized in Figure 5.6. The RNA (usually after the separation on oligo-dT cellulose) is fractionated by size using sucrose density gradient centrifugation. The RNA is applied to the top of a pre-poured gradient, and during centrifugation the larger molecules move down the centrifuge tube faster. The contents of the tube are then fractionated (usually by piercing

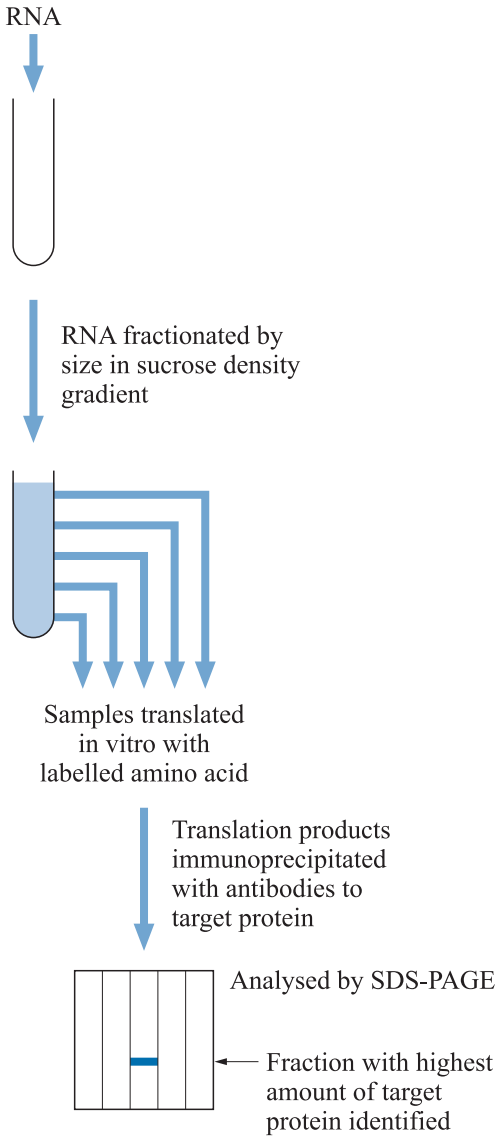


Fig 5.6 RNA fractionation. The figure shows a typical procedure to fractionate RNA by size and identify the fraction containing highest amount of RNA for the target protein.

the bottom and collecting individual drops or volumes). Each fraction will contain a different size-class of mRNA. A sample of each fraction is then translated in vitro. This is done either in an extract of wheat germ or in a lysate of rabbit reticulocyte cells. Both these preparations contain the necessary ribosomes, tRNAs and other components to translate the added mRNA with a low background of translation products from endogenous mRNA. At least one radioactively labelled amino acid is added with the mRNA, so that all the polypeptides subsequently synthesized will be radioactively labelled (and, therefore, can be detected in very low quantities, and also distinguished from all the polypeptides present initially in the lysate or extract).

The polypeptides produced after addition of the mRNA are then analysed with the antibodies. The latter are added to each reaction tube and will bind to the corresponding protein wherever it has been produced. The antibodies and any bound protein can then be precipitated and recovered. Rather than relying on simple precipitation of the antibody–antigen complex, it is wise to add something that will bind antibodies but which can easily be precipitated. Protein A-Sepharose is convenient for this. Protein A occurs on the outside of *Staphylococcus aureus* cells and binds IgG antibodies. It will, therefore, bind to the antibodies added to the translation products, and the antibodies will in turn be bound to the protein of interest (in any tubes where it has been produced). Because the Protein A is also attached to Sepharose beads (covalently coupled using cyanogen bromide), it can be pelletted and collected by centrifugation along with the antibodies and any bound antigen. (As an alternative to Protein A-Sepharose, intact *S. aureus* cells can be used.) The Protein A-Sepharose pellet recovered from each tube is then denatured and electrophoresed in an sodium dodecyl sulphate (SDS)–polyacrylamide gel and the location and amount of radioactive polypeptides determined. RNA fractions highly enriched for the mRNA of interest are identified as those from which large amounts of radioactive protein were precipitated by the antibodies and Protein A. These fractions are then used for the subsequent cDNA cloning to generate a library enriched for the cDNAs for particular mRNAs. Note that such a library is *not* representative of the total mRNA population from that tissue type.

3. **Normalized libraries.** A library made from cDNA that was prepared directly from mRNA will have a large number of members representing abundant RNAs and few representing the rare RNAs. If the cDNA we are looking for corresponds to an abundant RNA, that will increase the chances of finding an appropriate clone when the library is screened. However, if the cDNA we are looking for corresponds to a rare RNA, then we will have to screen a large number of members of the library. It is possible to increase the representation of rare mRNAs in a library by a technique called normalization. This technique is also very useful in construction of cDNA libraries for random sequencing in EST sequencing projects, as it reduces the number of times the same abundant cDNAs are sequenced.

   The principle most widely used relies on the kinetics of hybridization. If a collection of dsDNA molecules is melted (i.e. the strands are separated) and then allowed to reanneal, the more abundant DNA molecules will anneal more rapidly than the less abundant ones (as there will be more complementary strands for the former in the population). So, if we take a collection of cDNA molecules, melt them, allow reannealing of some but not all of the sequences and then collect the molecules that are still single stranded, our collection will be enriched for the rare species.

This principle underlies a number of normalization methods. One approach is shown in Figure 5.7 and is as follows. A conventional cDNA library is first constructed using the RNaseH method for cDNA generation and cloning into a standard vector. The cDNA inserts are then amplified using primers that flank the cloning site of the vector, melted by heating and then allowed to reanneal. Before reannealing is complete, the DNA is passed down a column of hydroxyapatite, which binds more tightly to double-stranded nucleic acids than to single-stranded nucleic acids. The single-stranded material, therefore, passes through the column. This eluate is enriched for the less abundant sequences (as more abundant sequences will be more likely to have annealed in

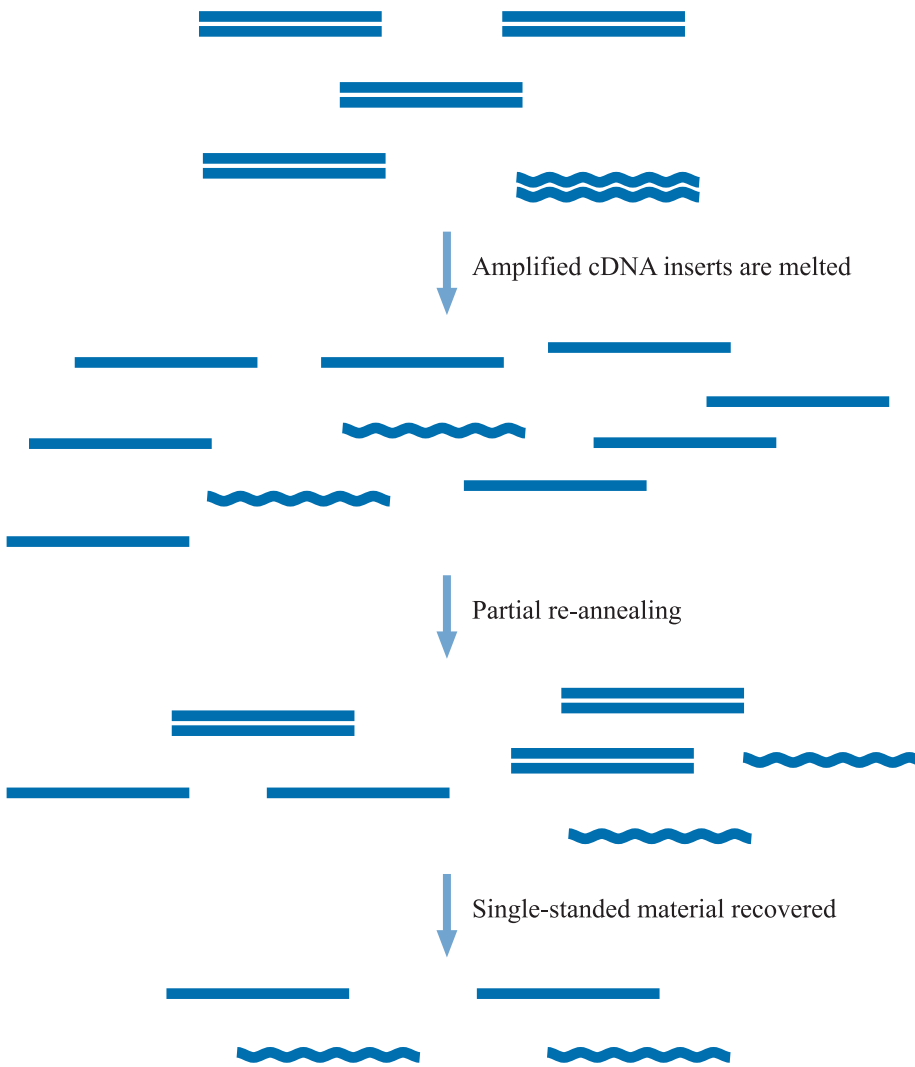


Fig 5.7 Normalization of a library. Solid lines represent an abundant sequence and wavy lines a rare sequence.

the time available). It can be enriched again by PCR if need be, and then cloned into a suitable vector as before.

5.4.3 Subtractive libraries

It is often very useful to make libraries that are enriched for sequences that are present in one sample but which are absent from another. These sequences might be present in the RNA from one tissue type but are absent from the RNA of another. Alternatively, they might be genomic DNA sequences that are present in a wild-type individual but are absent from a mutant that has a deletion in that region of the genome. We will look at the general principle first (shown in Figure 5.8), and then see how it can be applied. The nucleic acid from the cell type that contains the sequence we are interested in is called the tracer. The nucleic acid from the cell type lacking the sequence is called the driver. Tracer and driver are prepared in single-stranded form and mixed, with the driver in stoichiometric excess (usually at least 10-fold) over the tracer. The mixed sequences are allowed to hybridize.

(i) The driver can hybridize to complementary driver strands or to tracer.

(ii) The sequences in the tracer that are also present in the driver (the sequences we are not interested in) can hybridize to driver or to tracer. Because the driver is in excess, they are more likely to hybridize to driver.

(iii) The sequences in the tracer that are *not* represented in the driver (the sequences we *are* interested in) can hybridize to tracer or may not hybridize at all (depending on how the experiment is designed).

We clone specifically those sequences that have *not* hybridized to driver. This population of sequences is enriched for the ones that are present only in the tracer, and are the ones we are interested in. How do we arrange to clone only those sequences that have not hybridized to driver? There are various ways of doing this, three of which are given here.

1. **Sticky end compatibility.** Before the experiment, the tracer (which might be double-stranded cDNA or genomic DNA) is cut with an enzyme that gives sticky ends that are compatible with the cloning vector we will use. For cDNA, this might be *Eco*RI. The driver DNA is cut with an enzyme that generates small fragments with incompatible ends (say *Alu*I). The mix of hybridization products is added to the cut vector in a ligation reaction. The driver–driver hybrids will have ends that are incompatible with the vector, as will driver–tracer hybrids. Neither of these two kinds of hybrid, therefore, can be ligated to vector. The tracer–tracer hybrids will have compatible ends, however, and can be successfully ligated to the vector.

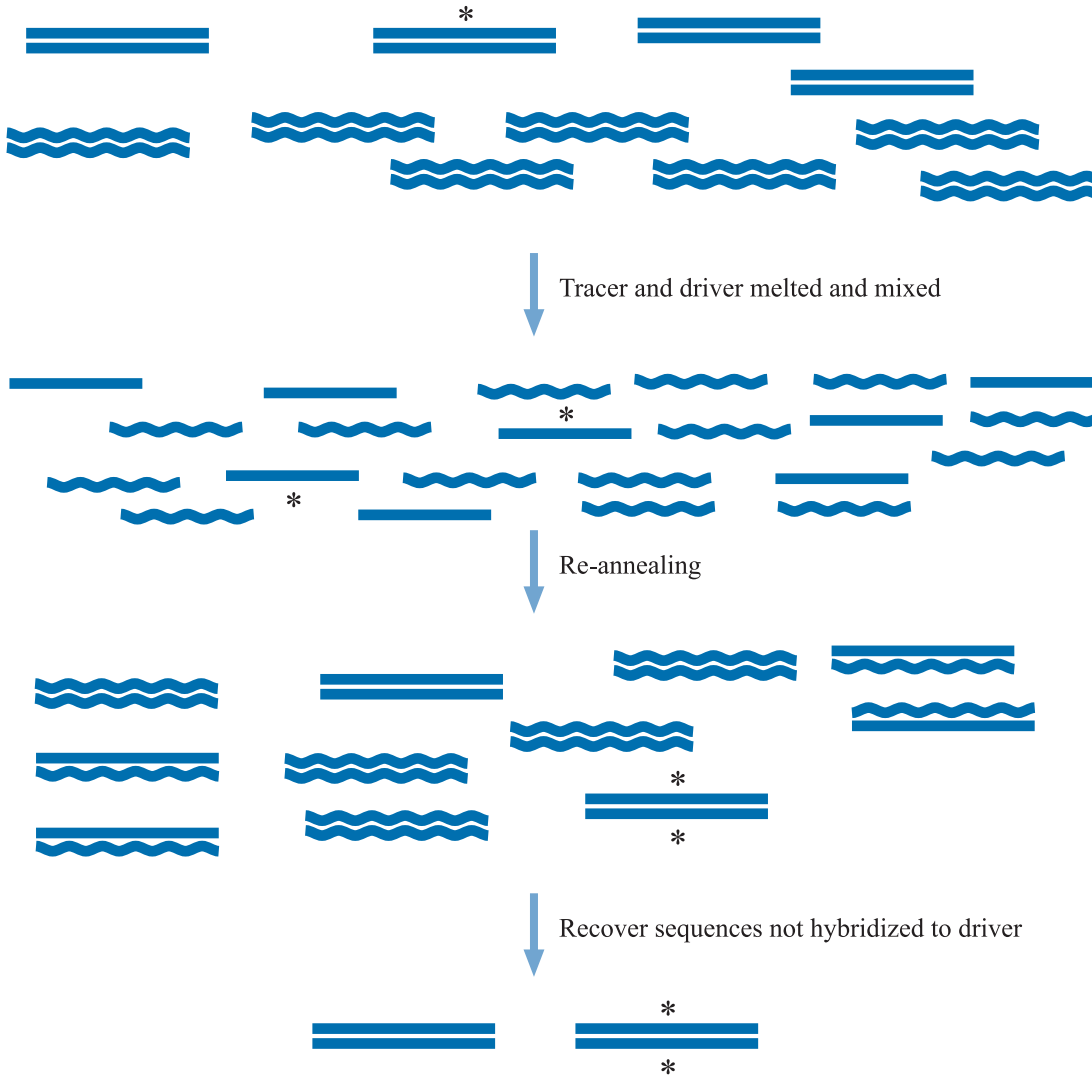


Fig 5.8 Subtractive libraries. Straight lines indicate tracer sequence and wavy lines indicate driver (in stoichiometric excess). Asterisks indicate tracer sequence absent from the driver population.

2. **Availability of complementary strands.** In this case, the driver is the first-strand cDNA from cells that lack the mRNA we are interested in. The tracer is mRNA from cells containing our target molecule. After hybridization, therefore, the mRNA species we are interested in will remain single stranded. The other mRNAs will form double-stranded DNA:RNA hybrids with the tracer. We can remove double-stranded material by passing the mixture down a hydroxyapatite column. The mRNA that is not retained by the column can then be used for cDNA synthesis in the usual way.

3. **Physical removal.** In this approach, the driver DNA is labelled in a way that subsequently allows its physical removal. One approach is to amplify the driver DNA in a PCR reaction that includes a nucleotide labelled with biotin. After hybridization between driver and tracer, streptavidin is added. This protein binds tightly to the biotin present in the driver strands. Material bound to streptavidin is then removed, selectively removing the driver sequence and anything base-paired to it. One approach is to use phenol extraction. This removes proteinaceous material (which partitions into the phenol phase or remains at the interface), and thus the DNA attached to it. The tracer DNA that is not hybridized to driver DNA remains in the aqueous phase and can be used for cDNA synthesis in the usual way.

If we are planning to use a subtractive cloning procedure to see which sequences are specifically expressed under particular circumstances (for example, by going on to do random sequencing of clones generated by subtractive cloning), then it is likely to be easier to find out which sequences are expressed using microarrays – assuming they are available for the organism in question.

5.4.4 Chromosome sorting

It may be that a gene of interest has been mapped to a particular chromosome. A DNA library produced not from the entire genome, but from just that chromosome, therefore, will be enriched for the gene of interest. Separation of intact chromosomes can be achieved using PFGE (see Chapter 1), but for cloning purposes a fluorescence-activated sorter is often used. The chromosomes should be in the metaphase state, because they will then be highly condensed and not too difficult to handle. They are isolated by gentle lysis of dividing cells. They are then stained with a fluorescent dye such as ethidium bromide (or a combination of dyes) and passed through the sorter, as shown in Figure 5.9. (Particular chromosomes can also be labelled by hybridization with specific fluorescently labelled DNA probes.) Each chromosome is driven by an electric field through a laser beam, which causes the dye to fluoresce. The amount of fluorescence is dependent on the amount of dye bound to the chromosome, and this in turn depends upon the size of the chromosome. The amount of fluorescence is measured by a detector, and the sorter uses the measurements to identify particular chromosomes. Application of an electric field across deflection plates, through which the chromosomes pass next, allows individual chromosomes to be deflected and collected. Once enough material has been collected, it can be deproteinized and used to prepare a library for that chromosome.

5.4.5 Microdissection

If a gene has been mapped to a particular chromosome or to a region on it, then it may be possible to use microdissection to obtain

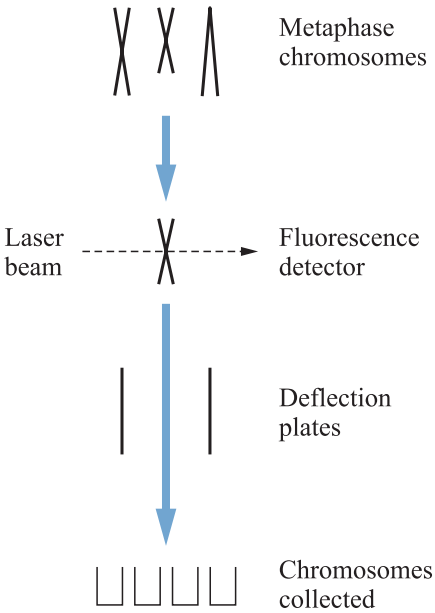


Fig 5.9 Chromosome sorting.

suitable material. Stained chromosomes are examined in a light microscope and whole chromosomes or appropriate regions removed with the aid of a micro-manipulator. The material is then collected in a pipette and used for cloning as before.

Chapter 6

Screening libraries

6.1 Introduction

So far, we have seen how libraries can be constructed using plasmids, phages and other vectors in *E. coli*. But it is not enough simply to be able to clone DNA at random. It is usually necessary to go on to identify members of a library that contain a piece of DNA with particular properties. Finding those members of the library is called screening. Most often, we screen libraries for sequences with a particular coding function to find the gene or cDNA for a particular protein. There are many different strategies for doing that, and we will look at those first. Sometimes libraries are screened for DNA with a particular function other than for coding, such as the ability to initiate or terminate transcription. We will look at those techniques too.

6.2 Database screening

One of the easiest ways to find a clone for a particular sequence (usually a gene or a cDNA) is to exploit genomic analyses. For many organisms, partial or complete genome sequences are available electronically in genomic databases. There are also EST databases, which are sequences of large collections of cDNA clones, typically sequenced at random from libraries constructed from different tissue types or under different conditions. The clones used to build the genomic and EST databases are generally widely available. If such databases exist for the organism you are interested in, then it is a simple matter to search the database by computer. This kind of screening is often described as in silico. If you can find the sequence you are interested in somewhere in the database, then you can obtain the relevant clones from a stock centre. How you search the database depends very much on the information available to you, as with the direct experimental screening of libraries. For example, we may have some sequence-related information for a protein of interest. (This information might be N-terminal sequence data or mass

spectrometric analysis of the sizes of proteolytic fragments.) The database is then searched for nucleotide sequences able to encode a protein with those properties. Another possibility is that we have some nucleotide sequence data, and wish to obtain a clone for a larger region. For example, we might have some cDNA sequence data and want to obtain a genomic clone. We then search the genomic sequence database for sequences matching our cDNA. A third common possibility is that we have sequence data for a homologue (perhaps the gene or cDNA we are interested in from a different organism). We then search the database for sequences similar, but not necessarily identical, to the sequence we already have. There are many programs available for this kind of analysis; one of the most common is BLAST (Basic Local Alignment Search Tool), which has a range of options for searching with or for nucleotide or protein sequences. Many genomic and EST databases are well described, in that the locations of potential coding regions have already been predicted from the sequence data, and possible functions of the proteins encoded suggested on the basis of similarity to other proteins of known function. This description of the database accessions is referred to as the annotation. It may be sufficient to search the annotation information in the database for terms of interest, rather than searching the actual database sequences themselves.

6.3 Experimental screening for coding function

6.3.1 Principles

We will assume throughout this section that the aim is to find library members encoding a particular protein. Screening for sequences encoding RNAs, such as rRNAs, is also possible using similar approaches. The techniques that are commonly used can be divided broadly into two classes:

(a) those that rely on nucleic acid hybridization;
(b) those that rely on expression of the coding function in vivo.

When screening very small libraries, methods relying on expression of the coding function in vitro may be useful. The actual approach taken, therefore, will depend on what kind of library is to be screened and what is available for the selection process (nucleic acid probes, antibodies, etc.). How to choose the right approach is discussed in more detail in Section 6.3.6. The techniques are most commonly used for screening libraries constructed in *E. coli* hosts, but they can generally also be used to screen libraries constructed in other hosts.

6.3.2 Screening by nucleic acid hybridization

Most of the nucleic acid hybridization screening techniques rely on one kind of experiment: the colony lift or plaque lift. Considerable variation is possible in the nature of the sequence used to probe the

library to find the right colony, though. We will look first at the basic experiment and then discuss the kinds of probe available. We will then look at three more specialized applications of this technique: chromosome walking, transposon tagging and breakpoint cloning.

The colony lift or plaque lift

This is sometimes called the Grunstein–Hogness technique, after the people who developed it. In essence, it is an experiment to find out which members of a library contain sequences complementary to a 'probe' DNA sequence. The library may be either *E. coli* colonies on a plate (in which case the experiment is called a colony lift) or plaques on a bacterial lawn (in which case it is called a plaque lift). The technique (Figure 6.1) is as follows:

(a) A piece of a suitable membrane (usually a form of nitrocellulose or nylon) is laid onto the plate for a few seconds and then peeled off. It will bring with it phage particles or bacterial cells adhering to the membrane. It is useful to carry out the experiment with duplicate membranes to distinguish genuine signals from random background ones (which can be caused by bacterial contamination).

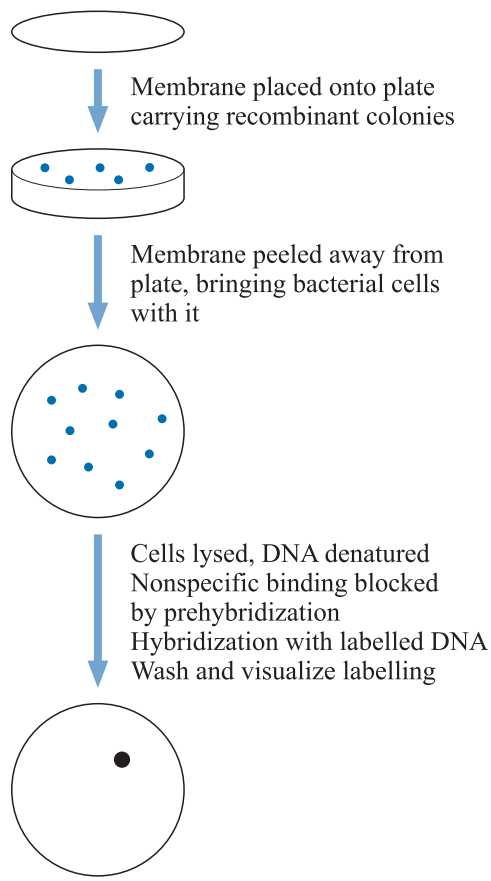


Fig 6.1 Colony lift. The method is described in the text. The same procedure can be used with a phage library; instead of plasmid DNA from bacterial colonies, phage DNA will be adsorbed onto the membrane.

Originally, nitrocellulose was the preferred membrane. This binds DNA very efficiently and, therefore, made the technique very sensitive. However, it has the disadvantages of being very brittle (and therefore difficult to handle without breakage) and also inflammable. Subsequently, nylon membranes were used; these are more durable than nitrocellulose, although they do not always offer the same sensitivity. Better sensitivity can be obtained using derivatized nylon membranes, which bind DNA more effectively. An alternative to transferring colonies to the membrane as described is to spot samples of cultures containing different members of the library directly onto the membrane in a regular array. Although this is time consuming to do manually, the procedure can be automated and is often included as part of large-scale genome analysis projects.

(b) The membrane is treated with sodium hydroxide to lyse bacterial cells and to denature phage proteins (in the case of a plaque lift) and the DNA.

(c) The membrane is treated with a neutralizing buffer, and baked at 80°C or irradiated with UV light to bind the DNA irreversibly to the membrane. The baking or cross-linking may be omitted with certain types of membrane.

(d) Non-specific DNA binding sites on the membrane are then blocked by treatment with a 'prehybridization' mix containing non-labelled DNA of non-specific sequence (often from salmon or herring testes or sperm). The membrane is washed and then labelled probe DNA is added and allowed to hybridize to any complementary sequences on the membrane.

(e) After hybridization, the membrane is washed again and the position of labelled probe DNA is determined. This is often done using chemiluminescence. There are various labelling methods available for this, but commonly the probe is labelled by incorporation of fluorescein-conjugated nucleotides. Detection is by an antifluorescein antibody conjugated to horseradish peroxidase. Using appropriate substrates, the horseradish peroxidase catalyses a reaction that results in the emission of light detected with a photographic film. Radioactive labelling of probes is also possible, with detection by autoradiography.

(f) The position of bound probe DNA indicates the position of plaques or colonies containing complementary, and therefore homologous, sequences.

The precise conditions for annealing of the probe and subsequent washing are critical. The important parameters are the size of the probe, the proportions of A's, G's, C's and T's, the ionic concentration of the hybridization buffer and the presence or absence of other agents, such as formamide, which destabilize base-pairing in DNA. These parameters determine the maximum temperature at which the probe and its target DNA (if fully complementary) will remain base-paired. Hybridization must, therefore, be carried out at

a temperature somewhat lower than this maximum. Carrying it out at an even lower temperature allows some mismatching between the probe and its target, so that the probe need not be a perfectly complementary sequence. This will be discussed later in this section.

The key to success in these experiments is to use a suitable probe. The following sections describe the types most commonly used.

1. **DNA probes**. We might have the following available:
   (a) PCR product. We might have available sufficient protein or DNA sequence information (either for the gene of interest itself or for a related one from the same or from a different species) to allow us to amplify a PCR product from genomic DNA or cDNA. Chapter 2 gives more information on primer design for PCR products. We could then use that to probe a library.
   (b) cDNA clone from a library. If a sequence has been isolated from a cDNA library (perhaps by random sequencing in an EST project) it can then be used to probe a genomic library to identify genomic clones. It can also be used to re-probe the same cDNA library (or a different one) to identify more cDNA clones.
   (c) Genomic DNA clone from a library. This can likewise be used to screen a cDNA library or a genomic library. Screening a genomic library allows the isolation of clones containing more copies of the gene (particularly useful when dealing with multigene families) or clones overlapping with the first (see the 'Chromosome walking' section later in this chapter).

   It is of course no use screening a library by hybridization with a clone containing the same vector sequence that was used to make the library, because the vector will hybridize to all the members of the library. Either the probe must be separated from the vector, perhaps by PCR amplification of the probe region or excision with an appropriate restriction enzyme followed by agarose gel electrophoresis and recovery from the gel, or the probe must be in a vector that does not have any sequence similarity with the library to be screened. It is notoriously difficult to separate DNA fragments from all possible contaminating DNA fragments in a gel, so use of a completely different vector would be more reliable. Unfortunately, so many vectors have sequences in common (such as particular drug resistance genes or origins of replication) that this is usually difficult, and so careful separation of probe from vector is necessary.

   DNA probes used for screening libraries do not have to match the target sequence completely. As a rough guide, the melting temperature of a perfectly matched probe is given by the following equation:

$$T_m = 81.5 + 16.6 \log_{10}[Na^+] + 0.41(\%GC) - 600/N$$

   where $[Na^+]$ is the molar concentration of sodium ions, %GC is the percentage of GC base-pairs in the probe and N is its length.

Hybridization and washing when using DNA probes are usually carried out at a few degrees below this melting temperature. A 1% mismatch between the target and probe sequences lowers the melting temperature by 1–1.5°C, so hybridization and washing at lower temperatures can be used to allow for mismatch. The greater the degree of mismatch allowed, the less stringent the hybridization is said to be, and the greater the possibility of the hybridization of the probe to the 'wrong' clones by chance complementarity.

The ability to allow for mismatch between probe and target means that they do not have to come from the same species. If they do come from the same species, the hybridization is said to be homologous. If not, it is said to be heterologous (even though there must be homology between probe and target for hybridization to work). So, once a sequence has been cloned from one species, it becomes much easier to obtain it from other species too.

2. **RNA probes**. Under some circumstances, RNA may be used as a probe. If we have a particular RNA available, then it can be used as a specific probe for the corresponding DNA species. Complete purification of individual RNAs is rarely possible, except for species such as rRNAs and tRNAs. However, RNAs can be synthesized in vitro, as discussed in Chapter 8, and these can be used to make probes.

   Populations of total RNA isolated from appropriate tissues can also be used. An example of this is the differential hybridization approach, sometimes called positive–negative screening. This offers a way of identifying clones for RNAs that are present in one population but not in another (e.g. as a result of treating cells with a stimulus that leads to the induction of certain genes). The colony or plaque lifts are first probed with labelled RNA from one population (say the non-induced cells). Once hybridizing colonies or plaques have been located, the labelled RNA is stripped from the membrane (usually by washing at a temperature greater than the melting temperature). The membrane is then probed with labelled RNA from the second population (in this case the induced cells) and the hybridizing colonies or plaques located. The colonies or plaques that hybridize in the second probing, but not in the first, correspond to sequences that are induced by whatever treatment was used. The approach can be made more sensitive by enriching the RNA from the second population (and also the library if necessary) for sequences that are absent from the first, as described in Chapter 5.
3. **Oligonucleotide probes**. If information is already available on the amino acid sequence of the protein in question (usually obtained by direct partial amino acid sequence determination of the purified protein), it may be more convenient to synthesize chemically an oligonucleotide that would encode that sequence and use it as a probe. However, because of the degeneracy of the genetic code, it will not be possible to deduce the corresponding nucleotide

sequence with certainty. Although (in the 'universal' code) there is only a single codon for methionine and a single codon for tryptophan, there are between two and six codons for each of the other amino acids. The best regions of amino acid sequence for constructing oligonucleotides, therefore, are those that are rich in methionine or tryptophan residues (with single codons), or, failing that, residues with only two codons. Most organisms have a preference for some codons over others for the same amino acid, so it may be possible to use this information to make an informed guess as to which of several choices is likely to be used. Otherwise it may be necessary to have a mixed site in the oligonucleotide. This means that a mixture of two or more nucleotides is used in a particular cycle of oligonucleotide synthesis, and different nucleotides will be put into different molecules. The result is a mixture of oligonucleotides with slightly different sequences. If enough mixed sites are included, one can be sure that the 'correct' sequence will be among those present. However, as more and more mixed sites are used, that correct sequence will come to represent a smaller and smaller fraction of the total number of sequences in the probe, and the signal generated on the autoradiograph after screening the library will get correspondingly weaker and weaker. Furthermore, the greater the number of incorrect sequences present, the more likely it is that at least one of them will be able to anneal to another sequence in the library that is unrelated to the one we want, giving rise to a consistent, but false, positive signal in the screening process. An alternative to using mixed sites is to incorporate a neutral base such as inosine. This can pair (although not necessarily strongly) with all of the standard bases. Designing oligonucleotides, therefore, requires a careful balance between informed guesses and mixed or neutral sites. It is a very similar problem to the design of oligonucleotide primers for PCR discussed in Chapter 2.

The temperature T_m at which a short oligonucleotide (containing 18 nucleotides or fewer) will melt from its target sequence is given by the following equation:

$$T_m = (\text{number of A's and T's}) \times 2^\circ\text{C} + (\text{number of C's and G's}) \times 4^\circ\text{C}$$

For longer molecules:

$$T_m = 81.5 + 16.6 \log_{10}[\text{Na}^+] + 0.41(\%\text{GC}) - 600/N$$

where $[\text{Na}^+]$ is the molar concentration of sodium ions, %GC is the percentage of GC base pairs in the probe and N is its length.

Hybridization is carried out at a few degrees below this temperature to ensure that probe and target can indeed hybridize. Mismatches can be allowed for by lowering the temperature still further. However, this also increases the risk of the probe annealing to a sequence other than the true target and generating a false positive signal. If enough amino acid sequence is available, then it may be possible to make two oligonucleotide probes

(each for a different part of the protein sequence) and accept as genuine positives in the screening only those clones that hybridize to both oligonucleotides. However, in that case it would probably be better to use the two oligonucleotides in PCR to generate a single probe that would be longer and more specific.

It may be wise to use a cDNA library rather than a genomic library when screening using oligonucleotide probes. Because cDNA libraries are enriched for coding sequences, fewer members need to be screened in order to find a suitable clone and, therefore, the risk of the probe hybridizing to a 'wrong' sequence is lower.

Chromosome walking

Successive rounds of screening of a genomic library with DNA probes can be used to assemble an ordered collection of clones located in a linear fashion down a chromosome. This is often called chromosome walking. It is useful for obtaining sequences that are close on a chromosome to one for which a clone is already available, such as a nearby gene or a microsatellite or restriction-fragment-length polymorphism marker. Starting with this clone, a genomic library is screened and hybridizing clones are picked out. These will contain sequences overlapping with the first sequence to a greater or lesser extent. The degree of overlap can be assessed by Southern blotting and hybridization. Clones selected in this way are used to screen the genomic library again. Hybridizing clones are picked out and their degree of overlap with the starting clone assessed. If there is little or no overlap, then they must be further from the starting clone than those picked out in the first round of selection, as shown in Figure 6.2. The cycle is repeated and, each time, clones that are further from the original can be picked out. It may be possible to assess cytologically how far the clones have spread from the starting point, by hybridization to chromosomes *in situ*. Obtaining a clone for a gene on the basis of its position on a genetic map is often called positional cloning.

With some organisms, it may be possible to use chromosomal inversions to 'jump' along a chromosome rather than

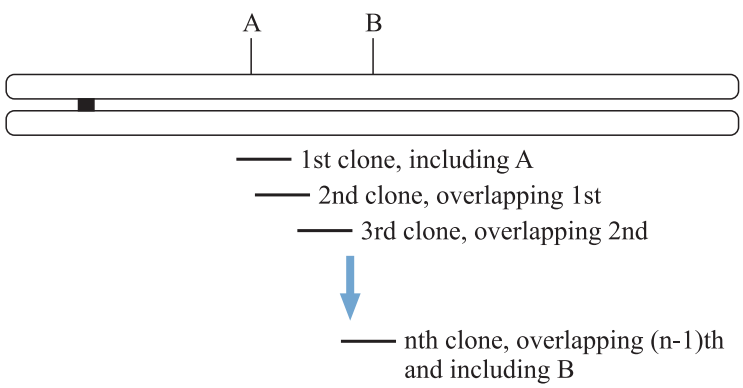


Fig 6.2 Chromosome walking. Starting from 'A' on the chromosome, a series of overlapping clones is picked out by hybridization. Ultimately, 'B' is reached.

simply walk. Suppose we have a clone covering the endpoint of a region that is inverted in some strains. Probing a genomic library from a strain carrying the inversion allows you to select sequences from the other end of the inversion, as shown in Figure 6.3. A modified form of jumping is also shown in Figure 6.3. Genomic DNA is cut into large pieces of 100–200 kbp or so and circularized by intramolecular ligation. The circles are then cut with a second enzyme. This will open them out to generate molecules that have been permuted. A library is then made from these molecules and clones hybridizing to the starting sequence selected. However, these clones will also contain material that is as much as 100–200 kbp away that has been brought adjacent by the circularization-opening process.

Transposon tagging

This approach is applicable when the organism under study contains transposons (transposable genetic elements) for which hybridization probes are available. These elements are DNA sequences that can move from one place in a chromosome to another, sometimes on the same chromosome and sometimes on a different one. This transposition may involve the duplication of the element, with a copy left behind at the original site. Transposons are found in many bacterial, animal and plant species. For transposon tagging to be feasible, it is not always necessary for transposons to occur naturally in the species in question. In some cases, they can be transferred from a species in which they do occur into one in which they do not by transformation, while retaining their ability to transpose. (Some species can efficiently integrate into their nuclear genome DNA that has no innate transposon ability at all. Tagging can be used in these species in the same way.)

For transposon tagging, we need to have a strain carrying a mutation caused by an insertion into the gene of interest. A genomic library is prepared from the mutant strain and screened with a probe for the transposable element. Clones that are picked out will contain the transposon and, flanking it, the gene of interest. If necessary, these clones can then be used to probe genomic libraries made from wild-type organisms to obtain the intact gene.

Transposon tagging can be complicated by the fact that there may be several copies of any given transposable element per genome, so it may not be possible to isolate the gene of interest unambiguously. Further screening of the clones selected may be necessary, using the other techniques outlined in this chapter. Sometimes a different strain has a mutation in the same gene, but caused by a different transposable element. This can be screened in exactly the same way, and the sequences that are picked out independently using both transposable elements are highly likely to be the ones of interest.

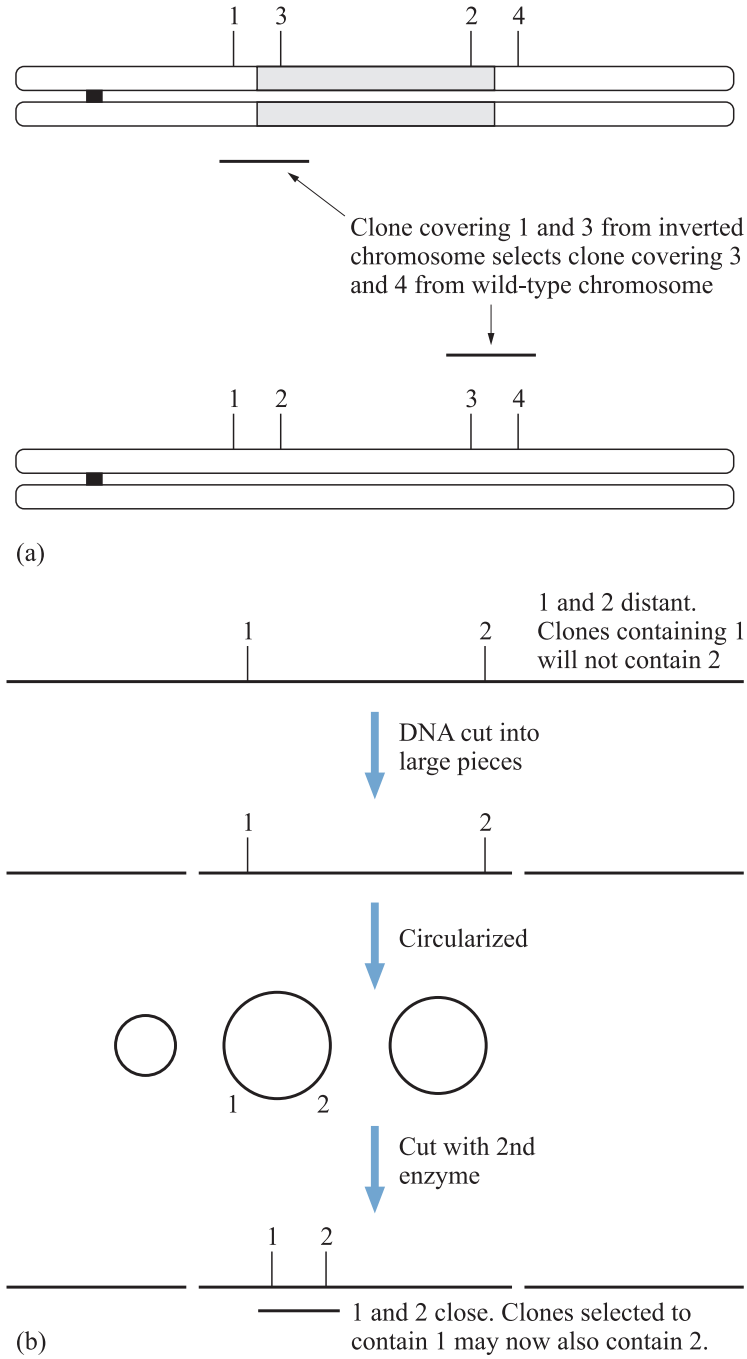


Fig 6.3 Two approaches to chromosome jumping. (a) The aim is to jump from point 1 to point 4. In the upper chromosome, the intermediate region has been inverted. A clone covering 1 from the inverted chromosome includes material (covering 3) that is adjacent to it, but which is adjacent to 4 in the non-inverted (lower) chromosome. The 1–3 clone can then be used to screen a library from the non-inverted chromosome to pick out 3–4 clones. (b) Regions of the genome are circularly permuted artificially by endonuclease digestion, circularization and redigestion. That brings sequences that were initially far apart (1 and 2) close together.

Breakpoint cloning

This approach can be used to identify the breakpoints of a reciprocal translocation, in which material has been exchanged between two chromosomes. These translocations are often identified as causing particular mutant phenotypes, and the ability to clone the sequences

disrupted by the translocations may help in identifying the genes that are affected. Clones are selected from a library (made with one of the translocated chromosomes) that contain material hybridizing to the wild-type forms of both chromosomes involved in the rearrangement. These clones are likely to span the breakpoint itself (as shown in Figure 6.4). This approach was used in a classic identification of the gene involved in Duchenne muscular dystrophy in humans, after the identification of a number of cases involving translocations between the X chromosome and chromosome 21.

6.3.3 Screening by expression in vivo

There are various methods that allow us to screen libraries for the protein product of the DNA sequence of interest, rather than the DNA sequence itself. These include direct selection for insert function and ligand binding by the expressed protein.

Direct selection for insert function

This approach is possible if the cloned gene can be expressed in the host and can complement a suitable mutant strain. For example, suppose the aim is to clone, using an *E. coli* host, the gene for the transport protein required for the uptake of a particular sugar by another bacterial species. The first requirement in this approach

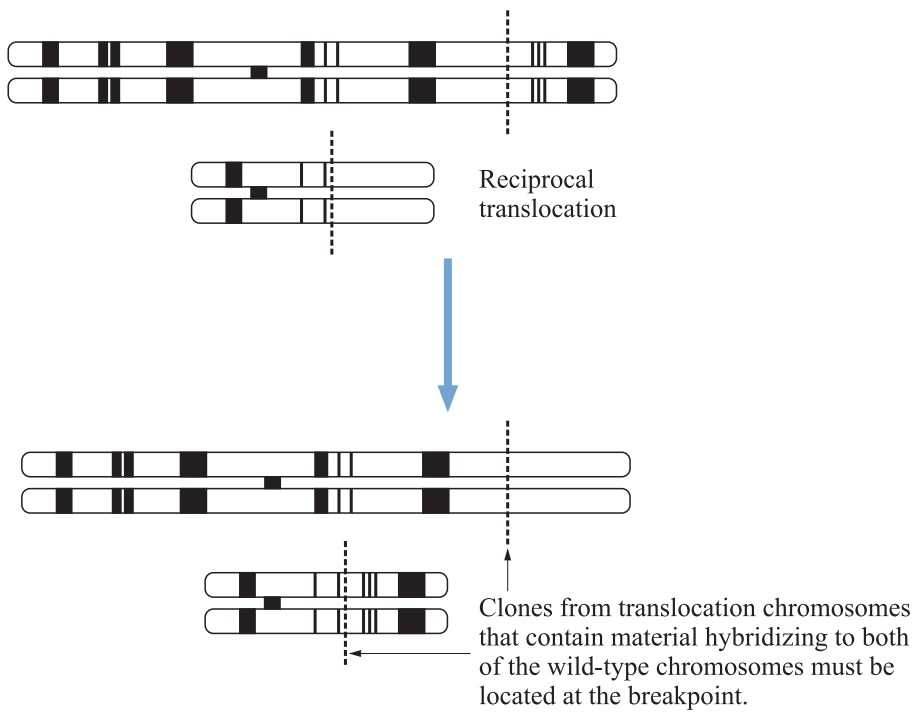


Fig 6.4 Breakpoint cloning. A reciprocal translocation has occurred between the upper two chromosomes to generate the lower ones. (Chromosome banding has been shown schematically to indicate this.)

would be to isolate by conventional microbiological techniques (or to obtain from a culture collection) an *E. coli* strain carrying a mutation in the equivalent gene, so that the strain cannot take up the sugar and, therefore, is unable to grow on medium containing the sugar as the sole carbon source. This strain will be used as host. A library is then made from the bacterial species we are studying and introduced into the mutant host strain by transformation, phage infection or whatever means is appropriate. After selection for the acquisition of recombinant molecules in the usual way, selection is then imposed for complementation of the host mutation, in this case for the ability to grow on medium containing the sugar as the sole carbon source. Individuals that are able to do that are likely to have acquired a wild-type copy of the relevant gene from the library.

This approach (summarized in Figure 6.5) need not be restricted to the identification of recombinants carrying nutritional markers. So long as a mutant strain can be obtained and complementation of the mutant selected, the approach can be tried. Note that the host species used for selection does not need to be the same as the source of the DNA in the library. However, the further apart in evolution that the source of DNA and the host species are, the less likely the technique is to work. There are a number of requirements and possible problems.

1. **Detailed understanding of the possible causes of complementation is needed.** Suppose (as is indeed often the case) there were a number of different proteins able to transport the sugar. A mutant that was deficient in transport could, therefore, be complemented

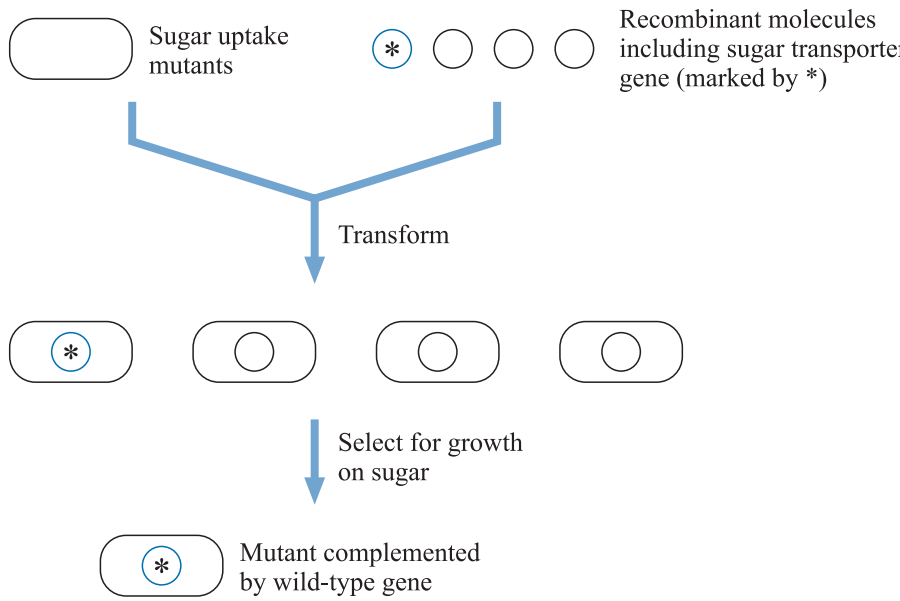


Fig 6.5 Selection for functional complementation. Asterisk indicates a gene for uptake of a particular sugar. In a sugar uptake mutant background, only the transformant carrying that gene can grow when the sugar is the sole source of carbon.

by any of a number of different genes from the library. Before using this approach, therefore, it is important to be aware of alternative means by which the initial mutant phenotype might be complemented.

2. **The mutant host must have a low reversion rate.** The proportion of recombinants that carry the gene of interest (and which, therefore, are able to complement the host mutation) is likely to be low, and may be lower than the spontaneous reversion rate of the mutation used. The consequence of this would be that most of the individuals selected would be the products not of complementation, but simply reversion of the original mutation and would not be of interest. However, the mutation used should affect one gene only, so mutant strains with large deletions (or polar mutations, which affect several genes) would not be suitable, even though they have a low reversion rate.
3. **Other features of the host genotype may be important.** It is not sufficient for the host just to contain a mutation in the gene of interest. The genetic background is also important. As with other hosts (Chapter 3), the one used here should be deficient in restriction enzymes, otherwise the incoming DNA may be degraded unless it has been protected by methylation. Ideally, the host should also be deficient in recombination, or the incoming DNA may become rearranged or integrated into the host chromosome (and would then be very difficult to recover). Many of the mutant strains used in more classical microbiological analysis do not have these characteristics, so they would need to be altered suitably. This may often be a relatively straightforward exercise in traditional bacterial genetics.
4. **The gene may not be expressed efficiently.** This is particularly likely to be the case if the gene comes from a very different species from the one in which the selection is carried out, because the sequences required for expression may well differ between the species. If the gene being screened for is prokaryotic, and even if it comes from the same species, it may still not be expressed if it is such a long way from its promoter that the promoter is unlikely to be present in the insert. cDNA sequences will be unlikely to have any of the associated expression sequences at all. It may be possible to circumvent some of these problems by the use of expression vectors (see Chapter 8), which are designed to allow the controlled expression of any genes cloned into them. If the gene being screened for is from a eukaryotic organism, then it might contain introns that were not excised after transcription in a prokaryotic host, so the approach would be unsuitable.
5. **If expressed, the protein may be non-functional.** This problem may arise in heterologous selection, where the library we are screening is from a very different species from the host. This is especially likely to be a problem if the protein has to interact physically with others for its function, because these interactions are less likely to work in a heterologous situation.

These difficulties may seem so great that the technique is unlikely to work, although very often only a partial restoration of function is needed to identify the right constructs satisfactorily. In fact, this approach does have two important advantages. One is the fact that there is no need to have prior information on DNA or amino acid sequence or antibodies to the product of the gene in question. The other advantage is that direct selection on a plate allows a very large number of recombinants to be screened quickly and easily.

Another approach to screening for insert function (although it is not strictly complementation) is to use cells such as *Xenopus* oocytes. DNA from collections of recombinants is transcribed in vitro and the transcription products are microinjected into oocytes. The RNA is translated within the oocytes, which are then screened for the function of interest. This approach has proved particularly useful for finding clones for membrane proteins with physiological functions that can be easily identified, such as the transport of ions or metabolites.

Ligand binding by the expressed protein

In this approach, the expression of a cloned gene or cDNA is detected not by the phenotype it confers, but by the ability of the protein it encodes to bind particular ligands. Very often, screening of libraries for ligand binding is done immunochemically, with antibodies. However, if the sequence we are screening for encodes a protein (such as a receptor) with a specific ligand, then we could use the ligand instead of antibodies. Or, if we are screening for sequences encoding proteins able to bind a specific DNA sequence, we could use that DNA. The procedure for immunochemical screening of plasmid libraries is summarized in Figure 6.6 and is somewhat similar to the colony lifts described earlier in this chapter.

(a) A membrane is placed onto the plate containing the colonies to be screened and then peeled off. This brings some of the bacteria from each colony with it. (The original plate can then be re-incubated to allow the colonies to regenerate.) The membrane is then exposed to chloroform vapour, to permeabilize the colonies, and then soaked in a lysis buffer. Proteins from the lysed colonies adhere to the membrane.

(b) After washing the membrane and blocking non-specific protein binding sites (usually with a suspension of dried milk powder), we incubate the membrane with the antibodies. If any members of the library are producing the relevant protein, then the protein will be present on the membrane at the position corresponding to those members of the library, and the antibodies will bind to it.

(c) The membrane is again washed. Now we can detect the position of bound antibody. This can be done by incubation with a secondary antibody that will cross-react with the first (primary) antibody, and has been labelled in a way that allows its detection. Commonly, detection is through chemiluminescence,

and secondary antibodies labelled for this purpose are supplied commercially. They are often conjugated to horseradish peroxidase, which catalyses a light-emitting reaction in the presence of appropriate substrates (Figure 6.7).

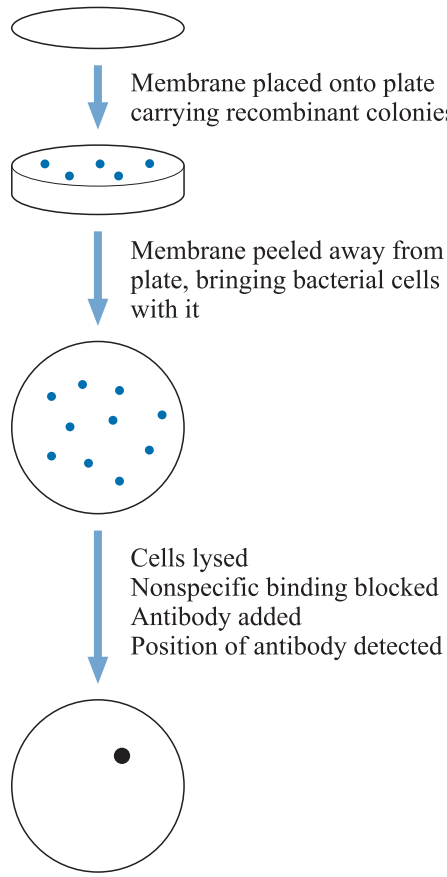


Fig 6.6 Immunochemical screening of plasmid libraries. If some other ligand is being used, then this would be substituted for antibodies in the penultimate step.

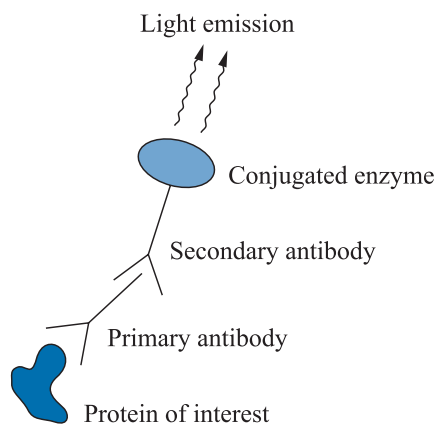


Fig 6.7 Detection of a protein using primary and secondary antibodies.

The sites where the primary antibody has been bound indicate the positions of recombinants expressing the gene for the protein of interest. (It is often a good idea to prepare duplicate membranes initially, so that spurious background signals can be distinguished from genuine signals that should be seen on both membranes.) The same general procedure can also be applied to screening phage libraries and to the use of other ligands. For the latter you might use a radioactively labelled ligand and detect its position autoradiographically.

The ligand-binding approach is very useful, although it requires the availability of antibodies or some other ligand. It is essential that the binding activity be preserved during the processing. It also requires that only a single protein species be necessary to bind to the ligand. If the ligand binds only to a heterodimer (i.e. a complex of two or more different proteins), then recombinants would have to produce all of the proteins to bind the ligand successfully; this is very unlikely. Screening by ligand binding also requires expression of the coding sequence in the host. This is particularly likely to be a problem when dealing with eukaryotic genes inserted into prokaryotic hosts. The expression signals are unlikely to be functional, and in the case of cDNA the signals are unlikely to be present anyway. The requirement for expression can be satisfied by the use of expression vectors, which are designed to give controlled expression of genes inserted into the vector. The presence of introns in the insert DNA is likely to pose severe problems, even if expression vectors are used, so this technique is largely restricted to screening cDNA libaries.

A variant of the ligand binding approach is called panning. Here, the ligand is immobilized on a solid-phase support and cells are passed over it. If the ligand-binding domain is exposed on the surface of any of the recombinants then they will be retained and can be recovered. This is particularly useful for screening libraries in cultured mammalian cells, where there is no cell wall to get in the way. Panning is not usually suitable for screening libraries in *E. coli*, as the protein product of interest is unlikely to be exposed on the outside of the cell.

A panning approach can be very useful with the phage display vectors described in Chapter 4, however. Phage display vectors contain foreign sequences inserted into the coding region for one of the phage coat proteins. For filamentous phage, gene III is often used. The coats of the recombinant phage, therefore, contain a hybrid protein, part of which is the product of the inserted sequence. A collection of recombinant phage is passed over a solid-phase support that has a ligand attached. Phage displaying a protein that binds to the ligand are retained and can be collected. The foreign DNA inserted into the coding region of a given selected phage must, therefore, encode a peptide sequence capable of binding to the ligand used. This method has been particularly useful for selecting sequences encoding fragments of antibodies to particular antigens. An extension of this method is ribosome display. A pool of ribosomes

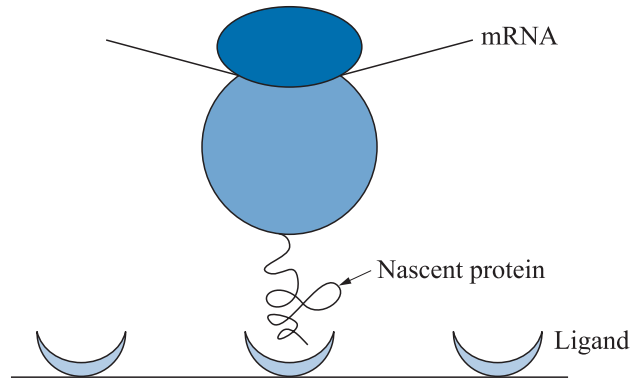


Fig 6.8 Screening by ribosome display.

that are in the process of translation (and thus contain mRNA and the corresponding nascent peptides) is passed over the solid phase with ligand attached. Ribosomes that are translating an RNA encoding a protein that binds the ligand are retained and the RNA recovered and cloned as cDNA (Figure 6.8). This method allows screening of much larger numbers of sequences than is convenient with a phage display library.

6.3.4 Expression of the coding sequences in vitro

Here, DNA constructs (usually plasmids) are isolated from the host and incubated in an extract capable of transcribing and translating them. This extract is often made from a lysate of *E. coli* cells, containing RNA polymerase, ribosomes, tRNAs, etc. Radiolabelled amino acids are used, so that any newly synthesized polypeptides can be readily detected. The products of the transcription–translation reactions are analysed by immunoprecipitation with antibodies to the protein in question, followed by electrophoresis of the immunoprecipitates in SDS–polyacrylamide gels and appropriate detection of the radioactive label. If a given construct is able to direct the synthesis of the relevant protein, then the latter will be precipitated by the antibodies and subsequently detected in the gel.

When translation takes place concomitantly with transcription, the processes are said to be coupled. It is also possible to carry out the reactions separately, in linked reactions. Here, the recombinants to be screened are incubated with a suitable RNA polymerase (which could be the *E. coli* enzyme or a more specialized one) to produce transcripts that are then translated in a separate reaction, normally by the use of a lysate from reticulocytes (immature red blood cells) or a wheat germ extract.

This approach relies on the ability of the extract to express the cloned sequences. Expression may well be easy for Gram-negative bacterial sequences (especially if the extract used for transcription and translation is from *E. coli*), but is much less likely for other sequences. Expression vectors may be helpful (see Chapter 8). The whole procedure from DNA isolation to SDS–polyacrylamide

gel electrophoresis is very time consuming, as well as requiring antibodies to the protein in question. The effort can be reduced by screening the DNA constructs in pooled batches. Once a pooled batch has been identified as containing a suitable recombinant, the individual members (or smaller pools) can be screened. The time-consuming nature of the approach means that it is useful only for screening small genomes, perhaps from viruses, organelles or large bacterial plasmids. Even in these cases, it is likely to be simpler just to sequence the whole genome of interest.

6.3.5 How do we confirm that we have the right clone?

Having isolated a cDNA or genomic clone, how do we confirm that it encodes a protein and, in particular, the protein that we are interested in? There are several features to look for.

1. **Open reading frames.** The easiest way to show that a piece of DNA could encode a protein is to demonstrate the presence of an open reading frame (ORF) in the nucleotide sequence. An ORF usually comprises an initiation codon followed by a long stretch without any termination codons in the same reading frame (unless the protein encoded is very small). If the gene contains introns, these will interrupt the ORF and may make it difficult to detect. This will not be a problem with cDNA. There are also computer programs that can identify sequences that are likely to be introns, by recognizing the splicing junctions or other features.
2. **Database searching.** We can use either the DNA sequence itself or the predicted protein product to screen nucleotide or protein sequence databases. For example, we may be able to show that the sequence we have cloned is similar to a previously determined sequence encoding a protein with similar function to the protein we are interested in. We may also be able to use computer searches of known and predicted protein structures to suggest what family of proteins the product of our cloned DNA sequence belongs to.
3. **Non-random base composition.** The requirement to encode an amino acid sequence imposes a non-random distribution of nucleotides in the DNA, which can be detected by computer analysis. Because introns do not usually have a coding function, this may help to distinguish them from exons, as the former will have a more random base composition.
4. **Function of the gene product.** If we know the biological activity of the protein we are interested in (if it is an enzyme, for example), then we may be able to assay for it in cells containing the cloned DNA.
5. **Distribution.** If we have neither an amino acid nor a DNA sequence for comparisons, nor an assay, things are more difficult. We could use northern analyses to check that the tissue or developmental distribution of mRNA from the gene we have cloned agrees with the distribution of the protein we are interested in.

At the protein level, we might be able to make antibodies to the protein that is encoded in our clone and use them to look at that protein's distribution. If it is the same as the distribution of the protein we are interested in, then this is further evidence that we have the right clone.

6. **Mutations**. Using the techniques described in Chapter 7, we could generate organisms with mutations in the gene we have cloned. If these also lack the protein we were interested in, or show an appropriate phenotype, this again suggests we have the right clone.

6.3.6 Choosing a strategy

For a probability P of having a particular sequence in a collection of recombinants, each of which has an insert equivalent to a fraction x of the genome, the collection must have $\ln(1-P)/\ln(1-x)$ recombinants. For example, to have a 99% chance of finding a sequence in a library of, say, 20 kbp fragments cloned from a genome of 3×10^6 kbp (i.e. with $x = 6.7 \times 10^{-6}$), about 700 000 recombinants would have to be screened. This is rather a lot to handle easily; it corresponds to a total amount of DNA over four times larger than the genome, and assumes that the screening process is fully reliable and that all copies of the sequence will be detected. Even for a 75% chance, over 200 000 recombinants would have to be screened, which is still a lot of recombinants. A cDNA library, however, is enriched for those sequences that are actually expressed, so a smaller number of recombinants need to be screened to obtain the sequence corresponding to a given protein. If the cDNA library was constructed using a tissue in which mRNA for the protein was particularly abundant, then the number of recombinants to be screened could be reduced still further. This would also be the case using one of the specialized libraries described in Chapter 5.

Exactly how a particular library is to be screened will depend on a number of factors, as well as the size of the library. These factors include availability of complete genome sequence data, the vector in which the library is constructed, the availability of antibodies, amino acid sequence, hybridization probes and mutants. There are no hard and fast rules for screening; the following are intended as guidelines for obtaining three types of genes. Chapter 10 gives some examples of strategies for screening and for subsequent manipulation.

1. **Genes from small genomes (e.g. viruses and organelles).** If the genome is small enough, then it may be easiest to sequence the entire genome and identify the gene of interest by searching the nucleotide sequence. The gene can then be amplified by PCR using primers derived from the known nucleotide sequence (or a related genome sequence if we do not have a genome sequence from the target organism). Alternatively, hybridization screening using probes from related organisms or based on amino acid sequence should be straightforward. The number of recombinants

to be screened will be small, so labour-intensive methods such as transcription and translation in vitro may be satisfactory, but they are still likely to be more time consuming than other methods. Mutant complementation is another possibility, depending on the availability of suitable strains.

2. **Bacterial genes.** It is likely to require too much work or expense to sequence a whole bacterial genome if we just want to find a single gene of interest. Very labour-intensive methods, such as transcription and translation in vitro, are also unlikely to be satisfactory, but otherwise the same principles apply as with smaller genomes. Given the range of complete bacterial genome sequences available, we may well be able to identify a homologue of the gene we are interested from the completely sequenced genome of a related bacterium and use that as a probe. To identify genes that are expressed under particular circumstances, simple sequencing of subtractive libraries may be appropriate. Alternatively, DNA microarrays may provide the same information.
3. **Eukaryotic genes.** Determination of a complete genome sequence to find a particular gene is even less likely to be feasible than with bacteria. However, sequence data from related organisms may allow PCR amplification or hybridization screening of genomic or cDNA libraries. Immunochemical screening of cDNA libraries is another possibility, depending on the availability of antibodies. In some circumstances, mutant complementation may be possible. Eukaryotes vary widely in the sizes of their genomes, and it may be easier to find the gene first in a 'model' organism with a small genome. It will certainly be easier to do that if there is a complete (or even partial) genome sequence for the model organism. For many organisms, there are EST databases containing sequence data for large collections of cDNAs, and these can be exploited in the same way as genomic sequence databases.

6.4 Screening for other functions: reporter genes

So far, we have concentrated on screening libraries for sequences that have particular coding functions. It is also possible to identify cloned sequences on other grounds, e.g. their ability to function as promoters, terminators, or origins of replication, or to encode proteins that interact physically with other previously characterized proteins. We will consider each of these in turn. Rather than screening the library directly for a promoter, terminator, or whatever, the methods we will look at rely on the sequence of interest modulating a gene whose expression level can easily be assessed. The latter gene is referred to as a reporter gene, as it reports on the activity of the sequence under study. We can also use reporter genes for assessing the activity of known promoters, terminators and so on. This is discussed in Chapter 8.

6.4.1 Promoters

A number of vectors exist for identifying sequences that are able to function as promoters in vivo. They are usually called promoter-probe vectors, and Figure 6.9 shows an example. It contains a reporter gene whose expression can readily be detected (in this case for chloramphenicol acetyltransferase, conferring chloramphenicol resistance) but whose promoter has been removed. The reporter gene is preceded by a cloning site. DNA is inserted into this cloning site and recombinants introduced into a suitable host. After selection for plasmid uptake (in this case by ampicillin resistance), transformants can be selected for chloramphenicol resistance. These transformants should have acquired a promoter in the cloning site directing transcription of the chloramphenicol resistance gene. (Alternatively, chloramphenicol acetyltransferase activity could be assayed to get a quantitative measure of promoter activity.) Note that the promoter-probe region is flanked in this vector by terminators, in this case from the *E. coli rrnB* operon. This is because very powerful promoters in the cloning site might interfere with other aspects of plasmid function, such as the ampicillin resistance.

There are many other markers, both selectable and non-selectable, that form the basis of promoter-probe vectors. Markers that are not directly selectable are generally designed to be readily detectable visually. Many are particularly widely used in screening libraries contained in hosts other than *E. coli*. The markers include the genes for green fluorescent protein, luciferase, beta-galactosidase and beta-glucuronidase. Green fluorescent protein (GFP) comes from a bioluminescent jellyfish, *Aequorea victoria*. This jellyfish uses a protein, aequorin, to produce blue light (in response to binding of calcium ions). The blue light is absorbed by GFP, which then emits green light by fluorescence. If the cloned GFP gene is expressed artificially in an organism as a reporter, then the protein can be detected by excitation with blue or near-UV light. Variants of GFP have been developed, by modification of the coding sequence, which fluoresce at different wavelengths. These include red, yellow and cyan

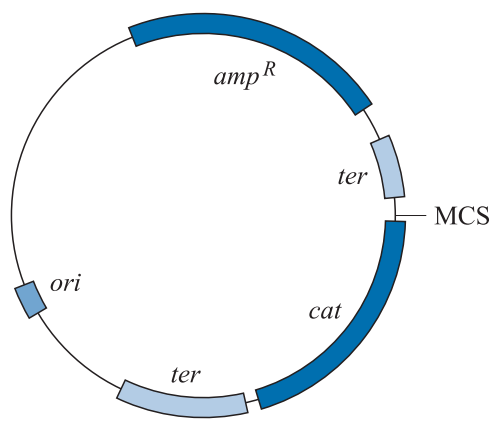


Fig 6.9 Promoter-probe vector pKK232–8 (5.1 kb). The promoterless reporter gene (*cat*) is for chloramphenicol acetyltransferase. Sequences for promoter screening are inserted into the multiple cloning site MCS. The *ter* sequences are transcription terminators. The amp^R gene is for ampicillin resistance and *ori* is the origin of replication.

fluorescent proteins. Fluorescent proteins from other organisms have also been exploited.

There are a number of luciferase genes available. The most commonly used ones are from fireflies and from luminescent bacteria, but a gene from the sea pansy *Renilla* is also used. Detection of luciferase requires the provision of the enzyme's substrate, which is used to produce light. Detection of beta-galactosidase and beta-glucuronidase requires the chromogenic substrates X-gal and its glucuronide equivalent X-gluc, respectively. The enzymes hydrolyse the substrates releasing the blue pigment. So luciferase, beta-galactosidase and beta-glucuronidase all require the provision of a suitable substrate for the reporter gene, ideally within the cell. If the cell is not readily permeable to the substrate, then this will be difficult. The fluorescent proteins, therefore, have the considerable advantage of not requiring the provision of a chemical substrate.

6.4.2 Limitations

Caution should be used when interpreting the results of experiments using promoter-probe vectors. The characteristics of promoters vary widely from species to species (both in sequence and in the proteins that bind to them), and it is quite possible that a sequence that is a promoter in one organism will not function as a promoter in *E. coli* and vice versa. So, although *E. coli* can be used as a host for vectors to probe for promoters from other species, it is better to use as a host the same species (or one that is as closely related to it as possible) as the one whose DNA is being screened. There are promoter-probe vectors for a wide range of organisms. They all work on the same principle of determining the expression of a promoterless reporter gene from a piece of inserted DNA, and use the same reporter genes as those already described. The more closely related the source of the insert DNA to the host, the more reliable the promoter-probe vectors will be. Where the host species and target DNA species are the same, we describe the screening system as homologous. There is no certainty of reliability even with homologous systems, as very many promoters require a control signal for activation. If this control signal is not supplied when the selection is carried out, then promoters responsive to it are unlikely to be detected. A similar approach is often used to determine the effects of sequence modification on known promoters that have already been cloned. The modified promoters are introduced into the host organism and the activity of the reporter gene monitored.

6.4.3 Transcription terminators

Terminator probe vectors work on a similar principle to that of promoter-probe vectors. Putative terminators are inserted between a promoter and a reporter gene. Terminator function can then be assayed by measuring the reduction in reporter gene expression. Direct selection for terminator function is also possible. Figure 6.10 shows an

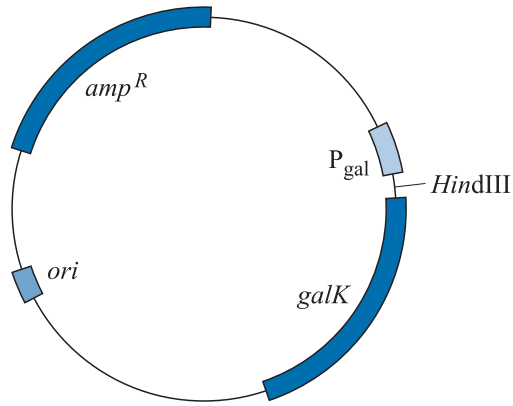


Fig 6.10 Terminator-probe vector pKGI800 (4.9 kb). P_{gal} is a promoter for *galK*, expression of which is lethal in a *galT*$^{-}$ background in the presence of galactose. The *amp*^{R} gene is for ampicillin resistance and *ori* is the origin of replication.

example for *E. coli* using the *galK* gene. Expression of this in a cell that is mutant for *galT*, which encodes galactose-1-phosphate uridylyl transferase (the next step in galactose metabolism) is lethal because of the build-up of galactose-1-phosphate. So the terminator probe vector is used in a *galK*$^{-}$ *galT*$^{-}$ host. If there is no terminator between the promoter and the plasmid *galK*, the plasmid *galK* gene is expressed, and the cell is phenotypically *galT*$^{-}$ and killed by galactose. If there is a terminator, the cell will be *galK*$^{-}$ *galT*$^{-}$ and, therefore, resistant to galactose. This allows the selection of sequences that have terminator activity in *E. coli*, but only under the conditions of growth on an agar plate. That does not necessarily give a conclusive indication of terminator activity under different conditions or in a different organism.

6.4.4 Origins of replication (autonomously replicating sequences)

It is also possible to select for sequences that can function as origins of replication. These are sometimes called autonomously replicating sequence (ARS) elements. The first step in selecting for an ARS is to remove the physiological origin of replication from a plasmid by digestion with a suitable restriction enzyme. The DNA under study is then ligated in its place, and recombinants introduced into an appropriate host. Selection is then imposed for a selectable marker on the plasmid. The removal of the plasmid's origin of replication means that stable propagation in the host is no longer possible and the marker cannot be retained. If a particular DNA sequence inserted into the plasmid contains a sequence able to act as an origin of replication, stable propagation of the plasmid is possible and the marker can be selected and retained. The same problems of interpretation apply as with promoter- and terminator-probe vectors. A further problem is that in some cases the probe plasmid may integrate (by legitimate or illegitimate recombination) into the host chromosome during the course of selection. It will, therefore, be stably propagated as part of the host chromosome, even though it does not necessarily have its own ARS element, and will give rise to colonies

in the selection procedure. The likelihood that this will occur can be greatly reduced by the use of a recombination-deficient host (see Chapter 3).

6.4.5 Interacting proteins

A very widely used technique is to screen members of a library for sequences that interact with a protein for which a cloned sequence is already available. This kind of screening is very often carried out in a yeast host (see Chapter 9), but there are similar systems designed for use in other eukaryotes and in prokaryotes, so we will consider them here. This is referred to as two-hybrid screening (Figure 6.11). The technique exploits the fact that many transcriptional activators comprise two domains. These are a DNA-binding domain and an activation domain. The activator binds through its DNA-binding domain to a sequence upstream of the gene under its control, and the activation domain then stimulates transcription. The two-hybrid screening system was made possible with the recognition that transcription activation can still be achieved if the two domains are linked by other proteins, rather than being part of one single protein (Figure 6.11a and b). A sequence encoding the protein for which we wish to find interacting partners (often called the bait) is cloned adjacent to the coding region for a DNA-binding domain of one of these transcription factors (Figure 6.11c). This is an example of a fusion protein, as discussed in Chapter 8. Members of the library to be screened are cloned adjacent to the coding region for the transcription factor's activation domain (Figure 6.11c). If a member of the library encodes a protein (the prey) that can interact with the bait, the DNA-binding domain and the activation domain will be brought together, linked by the interacting bait and prey. This results in activation of the gene that the transcription factor usually controls, and we screen for this activation. If the bait and prey do not interact, then the transcription factor will not be reconstituted and its corresponding gene will not be activated (Figure 6.11d).

The best-known yeast system uses the GAL4 transcription factor, which regulates genes involved in galactose metabolism, by binding to upstream activation sequences (UASs). The region encoding the DNA-binding domain is fused to the sequence encoding the bait protein to form the first hybrid protein. The sequences encoding potential interacting proteins are ligated into a vector that generates a fusion to the sequence encoding the GAL4 activation domain, to form a library of second hybrid proteins.

The members of this library are introduced into a host strain that (i) expresses the first hybrid bait protein and (ii) has a GAL4 UAS inserted upstream of a reporter gene. This reporter gene usually encodes beta-galactosidase or an enzyme involved in biosynthesis of a nutrient such as histidine or leucine (the *HIS3* and *LEU2* genes, respectively). Activation of the beta-galactosidase reporter gene can be detected by the production of blue pigment in the presence

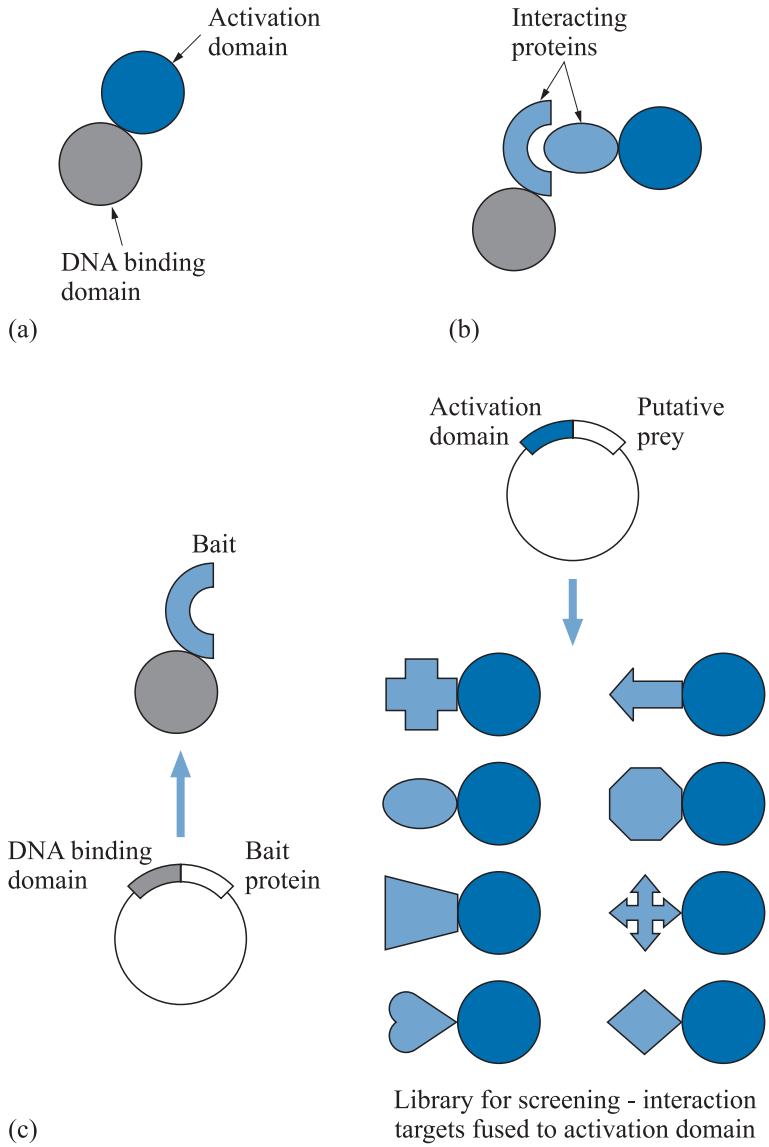


Fig 6.11 Two-hybrid screening. Panel (a) shows the usual configuration of the transcriptional activator used, with the DNA binding domain and the activation domain. Panel (b) shows the two domains linked by interacting proteins. Panel (c) shows a vector encoding a bait protein fused to a DNA-binding domain and the resulting fusion protein. It also shows a vector encoding an activation domain fused to a prey protein and the library of fusion proteins for screening. Panel (d) shows transcriptional activation where an interacting pair of proteins brings activation and DNA-binding domains together.

of X-gal. Activation of the *HIS3* and *LEU2* genes can be selected by the ability of cells that are otherwise *his3*$^{-}$ or *leu2*$^{-}$ to grow in the absence of histidine or leucine respectively in the growth medium.

In cells of the host strain that receive a library member encoding a hybrid protein that interacts with the bait hybrid protein, the activation and DNA-binding domains are brought together and the reporter gene is activated through the UAS. In cells receiving a library

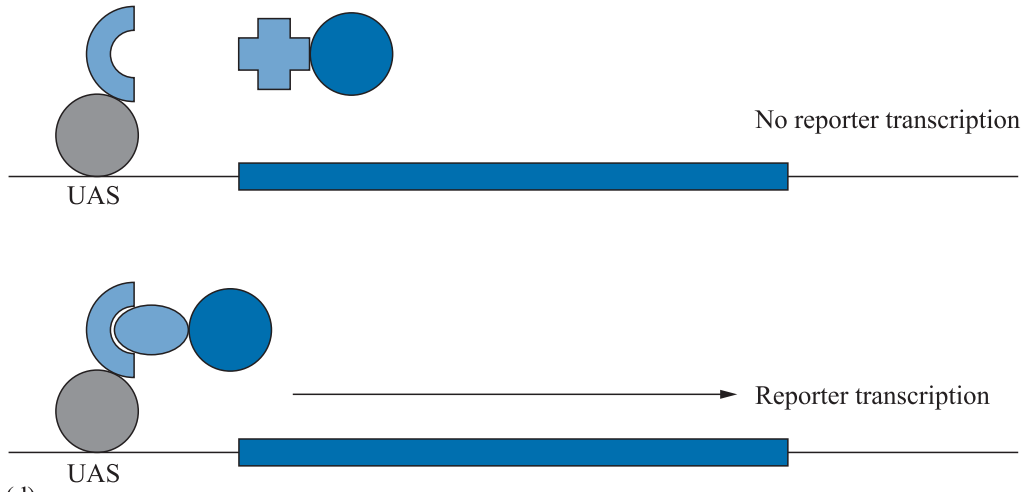


Fig. 6.11 (Cont.)

member that does not interact with the bait hybrid protein, the activation and DNA-binding domains are not brought together and the reporter gene is not activated. There are a number of potential limitations to this technique.

1. **The bait protein or some prey proteins may be non-specific activators of transcription.** If the former is the case, then cells expressing the bait without any prey may have the reporter gene activated. If the latter is the case, then some members of the prey library will activate the reporter gene even though they do not interact with the bait.
2. **Some bait or prey proteins may not be able to enter the cell nucleus.** Entry of both hybrid proteins is necessary for interaction to lead to reporter gene activation. If either or both cannot enter, then no reporter gene activation will be seen.
3. **Post-translational modification may be needed for the interaction.** If this is not achieved in the host strain used, then no reporter gene activation will be seen.
4. **The DNA-binding or activation domains may sterically hinder the interaction.** Again, this will stop the activation of the reporter gene.
5. **Non-specific interactions may occur between bait and prey.** This will lead to artefactual activation of the reporter gene.
6. **Transient interactions may not be detected.** Many very important protein–protein interactions, such as those involved in electron transfer reactions, are weak and not sufficiently stable for reporter gene expression to be activated.

There are a number of variations of the technique. In some yeast-based systems, the bacterial LexA protein is used to supply the DNA-binding domain instead of the GAL4 protein. There are also systems that can be used in bacteria and in mammalian cells. A development

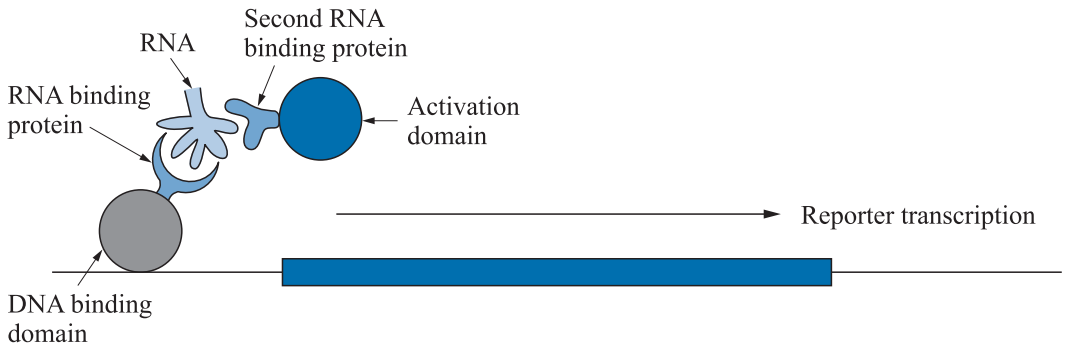


Fig 6.12 Three-hybrid screening. The figure shows reporter transcription where DNA-binding and activation domains are linked through a bait RNA-binding protein, its corresponding RNA and a second protein able to bind the same RNA.

of the two-hybrid screen is the three-hybrid screen. The latter has been used particularly to identify proteins that bind to RNA species or to other ligands. In the former case (see Figure 6.12) the bait construct has an RNA-binding protein fused to the usual DNA-binding domain. The host cell must also contain a plasmid that directs expression of the RNA of interest. The RNA-binding domain of the bait hybrid binds to this RNA. The prey hybrids are fusions to the activation domain in the usual way. Any prey hybrids that can also bind to the RNA will, therefore, result in the activation and DNA-binding domains being brought together, and the reporter gene activated in the usual way. In principle, the RNA could be replaced by anything that can bind both bait and prey. This might be another protein, or it may be a small molecule that is able to enter the host cells. This approach has been used to identify prey proteins able to bind to particular signalling ligands.

Chapter 7

Modification and mutagenesis

7.1 Introduction

So far, we have seen how to clone particular sequences and identify them. In Chapter 8 we will look at how these clones can be put to use directly at the DNA level or to direct the synthesis of RNA or protein. However, it is often the case that we need to modify sequences before using them. Here are just three of the many situations in which we may need to do this:

(a) We are trying to identify promoters and regulatory sequences, and need to make a mutation in a putative promoter or regulatory sequence to see whether that actually affects the efficiency of transcription initiation.

(b) We are interested in how the primary and higher order structures of a protein determine its function. It might, therefore, be necessary to modify the codon for an amino acid we believe to be at the active site of an enzyme and then assess the effects of that change on catalytic activity. Directed alteration of particular parts of proteins as a way of probing the relationship between structure and function or altering the function in a controlled way is often termed protein engineering.

(c) Genes are often cloned without our fully understanding the role that the proteins they encode have in the cell. Assessing that role may be possible by inactivating the endogenous gene in an organism to generate a mutant strain. This approach is often called reverse genetics, to emphasize the contrast with the traditional approach whereby a strain carrying a mutation with specific effects is characterized first and the relevant gene cloned and analysed subsequently.

Just what kind of mutation is needed, i.e. deletion, insertion, substitution and so on, will depend upon the particular biological problem being pursued. Having constructed a mutation, it is very important to verify by appropriate techniques (restriction mapping, DNA sequencing, etc.) that the right mutation has indeed been made,

and also that no others have been created unintentionally as a result of the techniques used. The approaches for making mutations can be divided broadly into two classes: those that rely on restriction enzymes and those that rely on oligonucleotide-directed DNA synthesis. We will deal with the former category first.

7.2 Methods based on restriction enzymes

We can exploit restriction enzymes, DNA polymerase and other enzymes to make a variety of changes. These include the removal or introduction of restriction sites, and the generation of larger insertions or deletions.

7.2.1 Removal of restriction sites

Sites can be removed by the procedure outlined in Figure 7.1. DNA is cut with the restriction enzyme whose site we wish to remove, to leave staggered ends. These are then rendered flush either by filling in with DNA polymerase or degradation with a suitable exonuclease, as described in Chapter 1. The ends are then religated. The net result is a small insertion or deletion that destroys the recognition site. Retreatment with the restriction enzyme will linearize any molecules

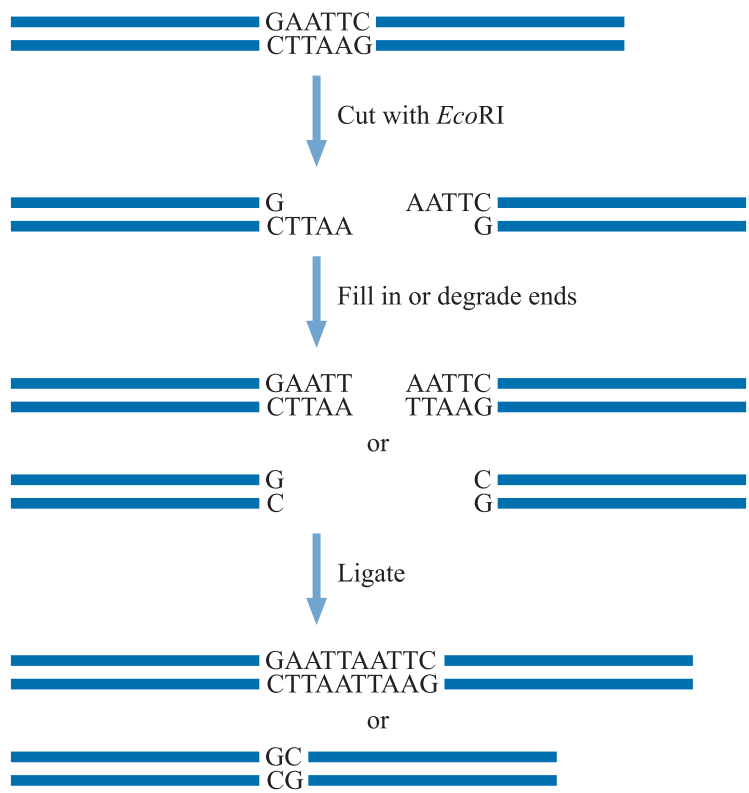


Fig 7.1 Deletion of a restriction site.

in which the site has not been successfully disrupted. A number of problems may be encountered in this approach.

1. **There may be multiple restriction sites.** There may be several target sites for the enzyme in the molecule. In this case, the initial digestion should be a partial one, so that molecules are on average cut at only one site. It will be necessary to map the molecules generated to find one with the right site(s) destroyed.
2. **The enzyme does not give staggered ends.** In this case, it is necessary to engineer a small deletion at the site using exonuclease, or a small insertion using a synthetic oligonucleotide.
3. **We may disrupt a functional sequence.** If the restriction site occurs in a functional sequence, such as a promoter or a protein-coding region, removal of the site may have undesirable consequences. This is particularly likely to be true where protein coding regions are concerned, since the procedure outlined in Figure 7.1 normally generates insertions or deletions of two or four base-pairs (as most sticky ends are of two or four bases). The procedure will, therefore, introduce a frameshift mutation into an open reading frame (since the sequence will be read in triplets). It may be possible to avoid this problem by introducing a slightly larger insertion or deletion. Alternatively, one could use oligonucleotide-directed mutagenesis to change the sequence at the restriction site without introducing a frameshift, as discussed later in this chapter.

7.2.2 Introduction of restriction sites

Additional restriction sites can be introduced near an existing restriction site by the insertion of a suitable chemically synthesized oligonucleotide carrying the appropriate sequence(s). An important example of this is the construction of the polylinker sequence in the pUC and M13 vectors. This was achieved initially by insertion of the oligonucleotide into an *Eco*RI site in the *lacZ* minigene. It is also possible to introduce restriction sites by oligonucleotide-directed mutagenesis of a region that contains a sequence similar, but not identical, to a restriction site.

7.2.3 Generation of insertions

Insertions are quite easily generated, usually by cutting at a suitable restriction site and then filling in the sticky ends with DNA polymerase, or ligating in a chemically synthesized oligonucleotide or a suitable restriction fragment from elsewhere.

7.2.4 Generation of deletions

Small deletions at a restriction site can be generated by cutting and degrading the single-stranded ends with an exonuclease. Alternatively, a complete restriction fragment can be excised and the molecule religated. Sets of unidirectional or bidirectional deletions can be generated as described in Chapter 1. Sometimes

a series of deleted regions (e.g. in various parts of a putative promoter) is replaced by some other DNA fragment of equal length, so as not to disrupt the spacing between other control regions, which may be important. This is called linker scanning.

7.3 Oligonucleotide-directed mutagenesis

The use of oligonucleotide-directed methods was an important development in this area. Before these methods were available, techniques exploited included the use of bisulphite ions to deaminate C residues in single-stranded regions to U residues, and error-prone DNA synthesis. However, these techniques were not very precise, in that they would routinely introduce multiple mutations, and they are now largely superseded. There are several slightly different methods in common use, although the underlying principle is the same. We can divide the methods into those that do not use PCR and those that do. We will look first at approaches that do not use PCR as they are a little simpler to understand, and then discuss methods that do use it. We will then look at methods of increasing the recovery of mutant sequences, in preference to wild-type sequences.

7.3.1 Mutagenesis without the use of PCR

This is summarized in Figure 7.2. We start with cloned DNA for the region to be mutated, which will be a template for DNA synthesis. We can use single-stranded DNA prepared using one of the filamentous phage vectors described in Chapter 4, or dsDNA prepared using a conventional vector. An oligonucleotide is synthesized, whose sequence contains the mutation we want to make but which is otherwise complementary to the template DNA. The oligonucleotide is allowed to anneal to the template. If the template is single stranded, this is straightforward. If the template is double stranded, the strands may first need to be separated by heating or by alkali denaturation. After annealing of the oligonucleotide, DNA polymerase and dNTPs are added. The oligonucleotide functions as a primer for the synthesis of the rest of the second strand. (Note that this requires the use of a polymerase that lacks 5′-3′ exonuclease activity. A polymerase that had 5′-3′ exonuclease activity would degrade the primer that carries the mutant sequence, once synthesis had proceeded all the way round the circular template, and replace the primer with DNA synthesized using the wild-type template.) The double-stranded molecule is covalently closed with ligase and contains a mismatch at the site of the mutation. It is then reintroduced into *E. coli*. Replication will generate two types of molecule: one with the wild type sequence and one with the mutated sequence. If mismatch repair takes place before replication, then the molecules will be all wild type or all mutant according to the direction of repair. If mismatch repair does not take place,

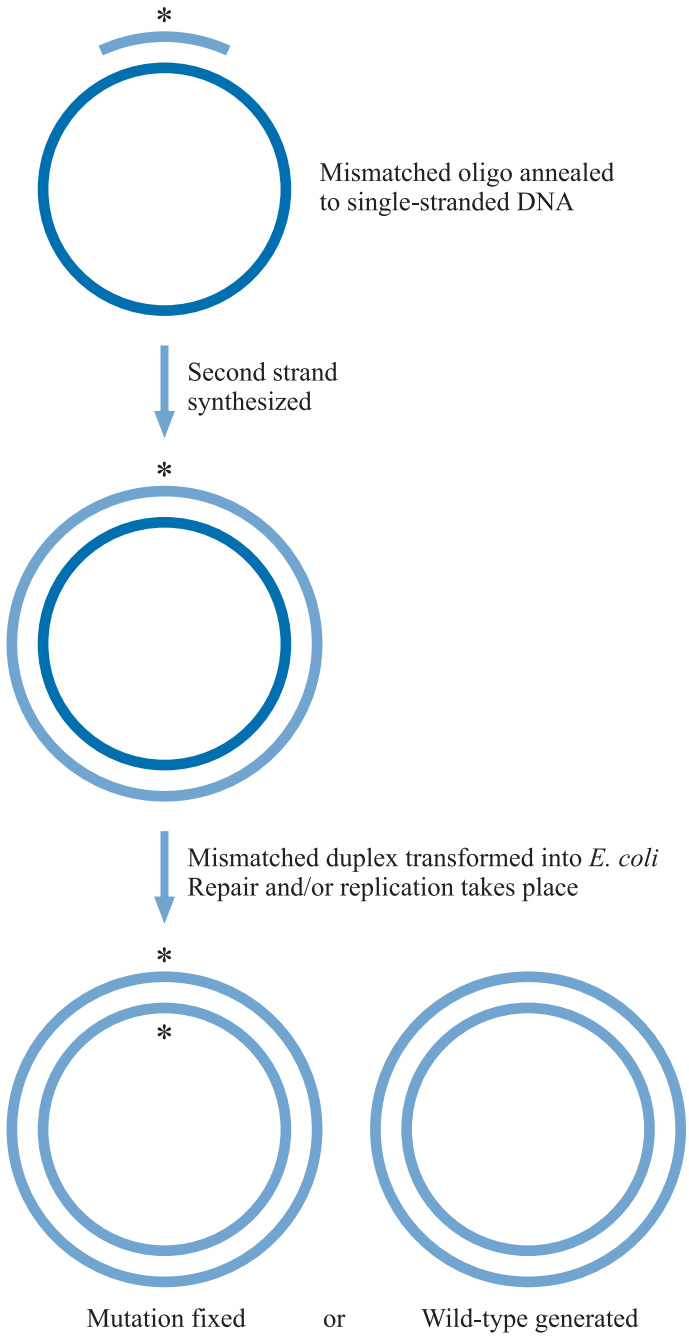


Fig 7.2 Oligonucleotide mutagenesis: the basic reaction. The mutation we need to make is indicated by the asterisk.

and a mixed population of molecules is likely to be present in each host cell, then the molecules must be separated by repeated streaking out of the host cells from single colonies or, if a phage vector is used, repeated plating out of the phage from individual plaques (termed plaque purification). DNA sequencing can then be used to check that the desired mutation has been incorporated and, importantly, that no additional unwanted mutations have been generated.

There are several modifications to this basic approach.

(a) More than one site can be mutated at a time, using an oligonucleotide with more than one mismatch to the target sequence.

(b) Several different mutations can be made *at the same site*, using an 'oligonucleotide' with a mixed site (i.e. a collection of oligonucleotides with different nucleotides at the position in question). DNA sequencing of putative mutant molecules re-isolated from the host can be used to verify which mutations have been generated. The use of mixed sites in an oligonucleotide to generate several mutations is often called cassette mutagenesis. It is possible to generate large numbers of mutations very quickly this way. For example, an oligonucleotide with two triply mixed sites could generate 15 different mutant sequences (six single mutations and nine double mutations).

(c) If the target sequence is from a large gene, then it may be easier to sub-clone a restriction fragment from it, mutate the fragment, verify the mutation by sequencing and then reassemble the gene by replacing the original wild-type restriction fragment with a mutated one. This reduces the chance of getting unwanted additional mutations at other sites in the target gene.

7.3.2 Mutagenesis using PCR

Rather than using a single round of synthesis of mutant DNA, we can use PCR to carry out multiple rounds of synthesis. As before, the principle is to use a primer that carries the required mutation. One approach is termed the 'megaprimer' method (shown in Figure 7.3). This is a two-stage approach and uses three oligonucleotide primers. Two of them are flanking primers, annealing at the ends of the target gene. The third anneals at the target site and contains the required mutation. PCR using this mutagenic primer and one of the flanking primers generates a product corresponding to part of the gene. This is the megaprimer. A second PCR is then carried out using the megaprimer and the second flanking primer. The product of this is the full-length gene with the mutation incorporated, and it can be cloned and sequenced in the usual way.

The second approach is more similar to the method we looked at in Section 7.3.1. It relies on PCR amplification of the cloned target sequence and uses double-stranded circular template DNA and two primers. These anneal at the target site, carry the required mutation and direct synthesis in opposite directions, as shown in Figure 7.4. After repeated synthesis (and ligation to seal nicks if required) the molecules are introduced into host cells. The majority of DNA is newly synthesized and contains the primers that carry the mutated sequence. DNA recovered after transformation of the host cells should, therefore, be more likely to have the required sequence.

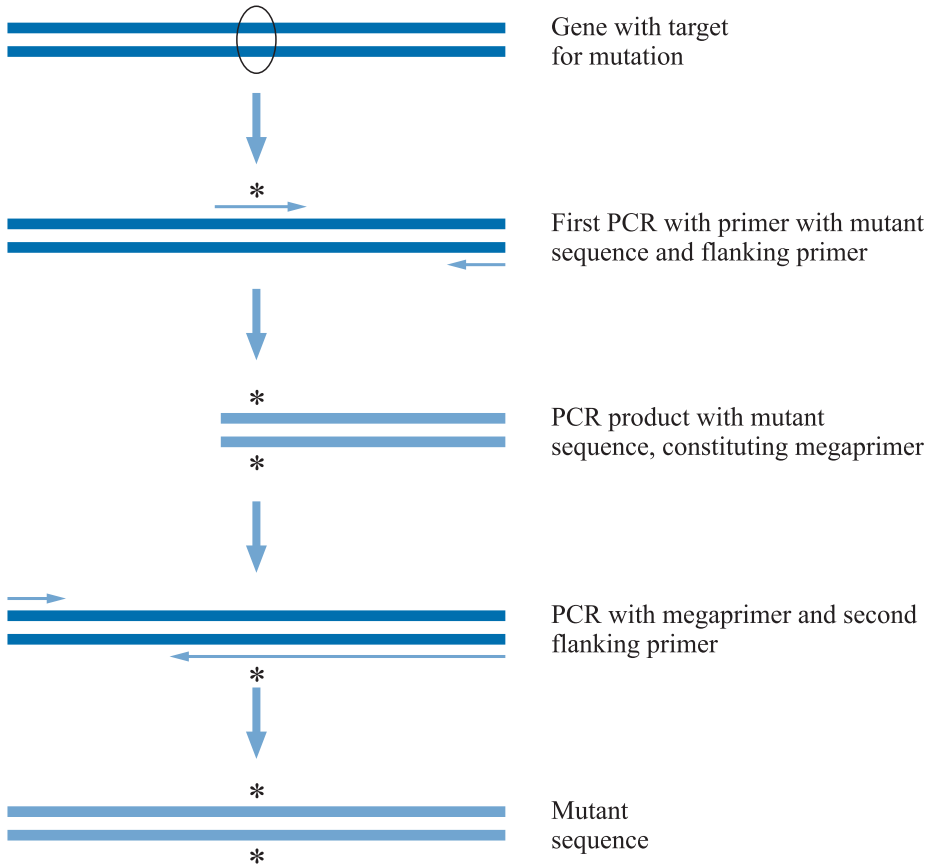


Fig 7.3 Mutagenesis using a megaprimer. For simplicity, only the product carrying the mutant sequence is shown. The asterisk indicates the mutation we need to make.

PCR is sometimes used to make DNA molecules with one or more random mutations, rather than molecules with specific mutations. This is called error-prone PCR. A PCR is carried out with primers that are fully matched to the template, but under conditions where there is a high error rate in DNA synthesis. This includes the use of elevated concentrations of magnesium and manganese ions, and unequal concentrations of dNTPs. An enzyme without proof-reading activity, such as Taq polymerase, is used.

7.3.3 Increasing the recovery of mutant sequences

In the approaches described in Sections 7.3.1 and 7.3.2, it would be expected that the majority of the molecules produced would be mutant. However, we often find empirically that we recover the wild-type sequence after transformation more often than expected. A number of factors may contribute to this bias in which molecules are recovered. Newly synthesized material made by the method shown in Figure 7.4 may still contain nicks and be susceptible to degradation in a host cell. Also, some molecules may be able to replicate faster than others, or they may affect host viability differently. Finally, mismatch repair might be biased in a particular direction. Therefore, the frequency with which mutants are obtained

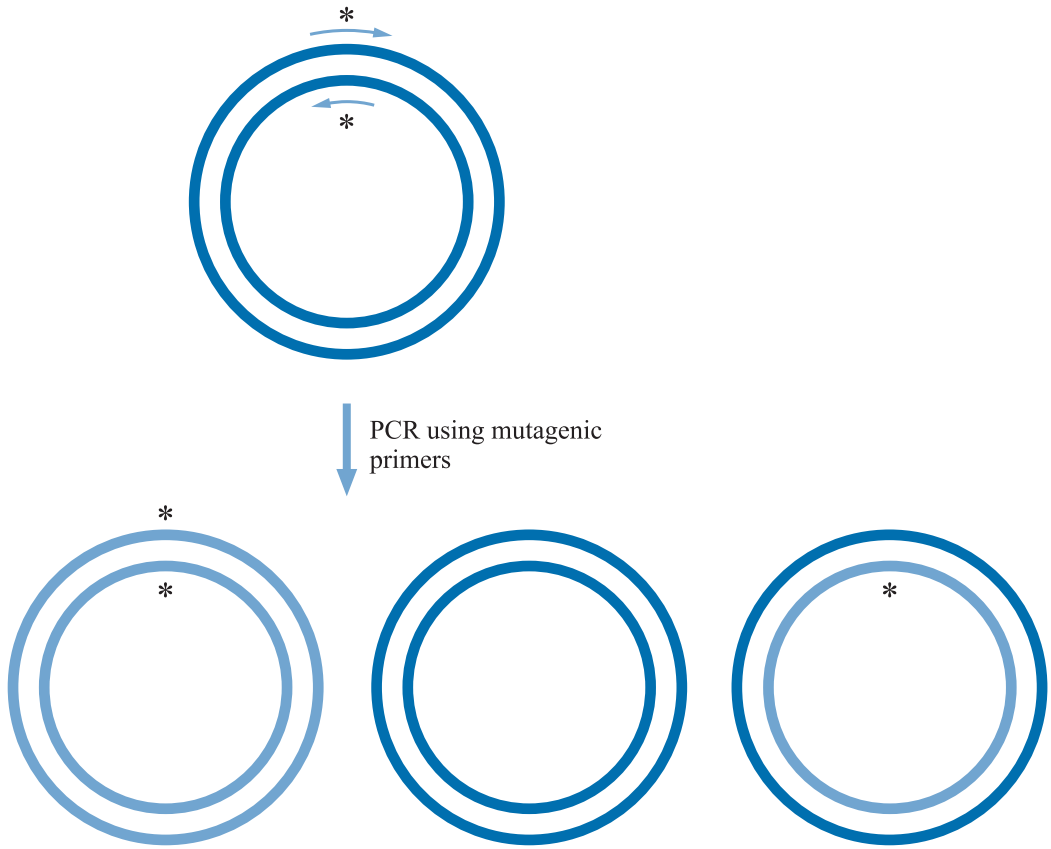


Fig 7.4 Mutagenesis using two mutagenic primers. Among the products are molecules carrying the desired mutation on both strands. Note that newly synthesized material may contain nicks.

may be so low that steps have to be taken to improve it. There are two approaches. One is to provide a selection for the mutant sequence and the other is to degrade the wild-type molecules. We will deal with the use of selection first.

1. **Selection for the mutant sequence.** This approach (Figure 7.5) is suitable for methods that are not PCR based. It uses a vector that contains (in addition to a conventional antibiotic resistance gene, say for ampicillin resistance) a gene for resistance to a second antibiotic, but which is non-functional because of a point mutation (*res* in Figure 7.5). In addition to the mutagenic primer to direct the mutation of interest, a second mutagenic oligonucleotide is used that will revert the mutant antibiotic resistance gene and make it functional. Both oligonucleotides are used simultaneously in the second-strand synthesis reaction. After second-strand synthesis and ligation to seal any nicks, the products are introduced into an *E. coli* host by transformation. This host carries the *mutS* mutation, so no mismatch repair takes place. Once transformants have been selected (in this case using ampicillin) they are selected for resistance to the second antibiotic. Plasmids in these transformants were derived from the second strand, synthesized using the mutagenic oligonucleotides. The plasmids should, therefore, also have the mutated sequence at the target site. (A similar procedure

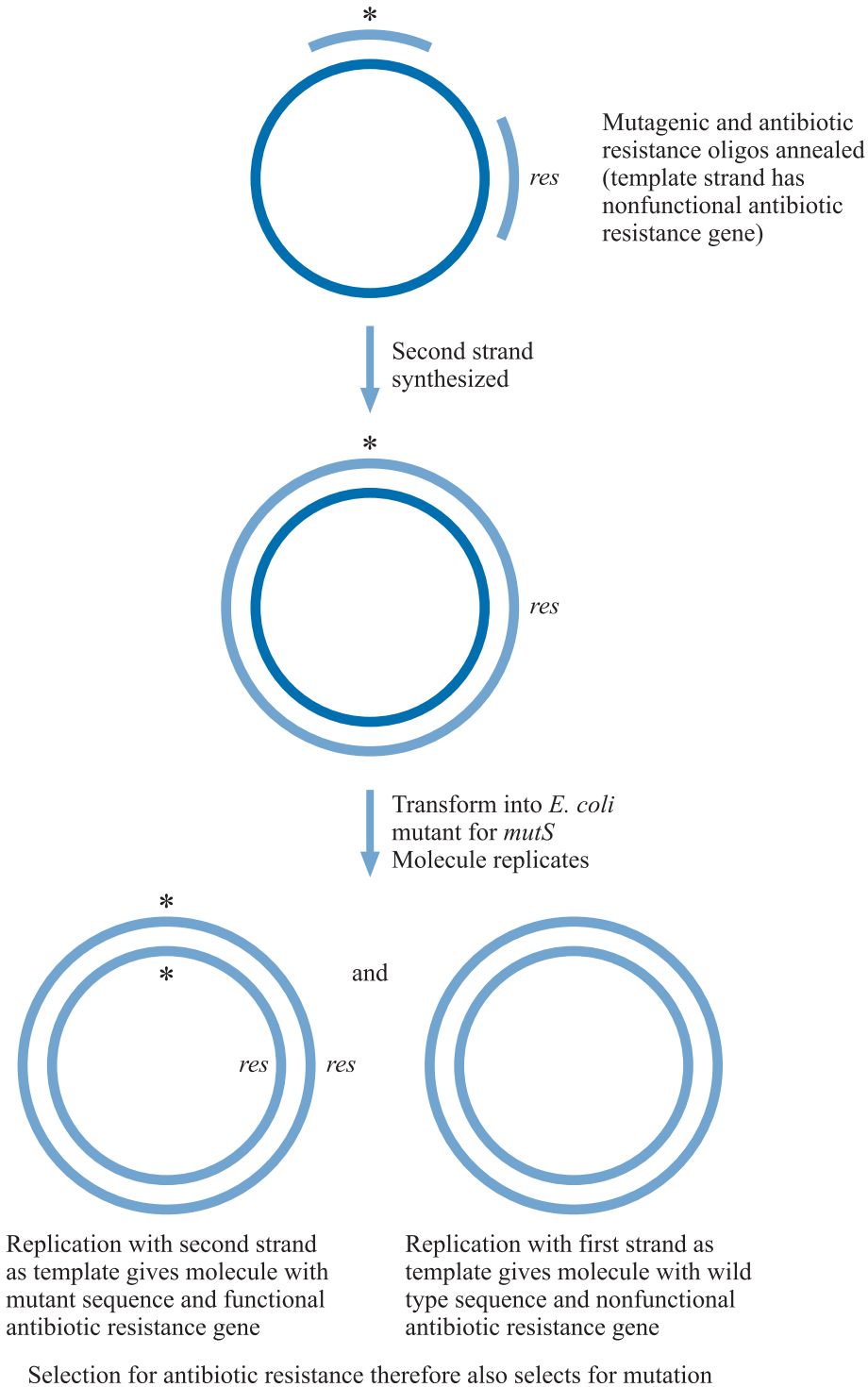


Fig 7.5 Oligonucleotide mutagenesis incorporating selection for mutant molecules. The sequence able to confer antibiotic resistance is indicated by *res*, and the asterisk indicates the mutation we need to make.

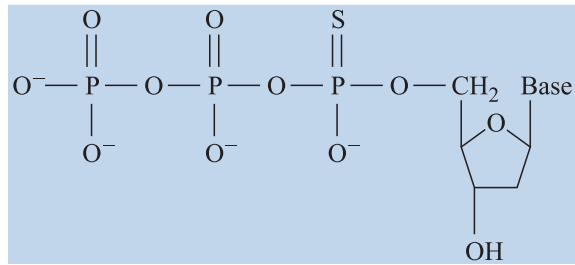


Fig 7.6 Structure of a phosphorothiorate nucleotide.

uses a mutant *lacZ* gene instead of the second antibiotic resistance gene, allowing chromogenic selection, rather than direct antibiotic selection.) Using two oligonucleotides in this way requires careful control of the annealing conditions, so that as many template molecules as possible have *both* oligonucleotides bound.

2. **Use of phosphorothiorate nucleotides to bias mismatch repair in vitro.** This approach is also suitable for use with simple second-strand synthesis methods. It uses a form of mismatch repair in vitro that is biased using phosphorothiorate nucleotide analogues. The general structure of these analogues is shown in Figure 7.6. The replacement of an oxygen atom by a sulphur atom in these nucleotides renders DNA containing such residues resistant to attack by a number of nucleases, including the restriction endonuclease *Nci*I and exonuclease III. This fact is exploited in the following way. The target DNA and mutagenic oligonucleotide are annealed as usual and the second-strand synthesis is then carried out using at least one phosphorothiorate nucleotide. We then treat the DNA with *Nci*I. This cuts the first strand, resulting in nicks, but it cannot cut the second strand because of the presence of phosphorothiorate nucleotides in that strand. The nicks in the first strand are then enlarged by treatment with exonuclease III. The first strand is thus largely degraded. It is then resynthesized by addition of DNA polymerase and dNTPs, using the second strand (which has the mutated sequence) as a template. Both strands will, therefore, come to have the mutated sequence. The resulting double-stranded material is then reintroduced into a host.
3. **Using template containing uracil to target wild-type molecules.** To understand this method we must first have some background information on how uracil can become incorporated into DNA and how cells deal with it. Uracil can be formed in DNA by the spontaneous deamination of cytosine, and in a subsequent round of replication the uracil would direct the incorporation of an A residue on the opposite strand where there had been G initially. The presence of uracil in DNA, therefore, is potentially mutagenic, and *E. coli* has a mechanism for its removal. This relies on the *ung* gene, encoding a uracil-*N*-glycosylase. This enzyme removes the uracil from the deoxyribose, leaving an apyrimidinic site

(i.e. a site in which the base is missing but the phosphodiester backbone is intact). The DNA strand containing the apyrimidinic site is then nicked by an endonuclease and the defective region is degraded and replaced by DNA polymerase I using the other strand as a template. Apart from the deamination of cytosine, U residues can also become incorporated into DNA by the use of dUTP rather than dTTP during replication by DNA polymerase. Cells normally contain only low levels of dUTP, which functions as an intermediate in the synthesis of dTTP. The level is kept low by the presence of a dUTPase. This hydrolyses the dUTP (catalysing the next step in the synthesis of dTTP). It is the product of the *dut* gene. U that is incorporated into DNA from dUTP will normally be removed in the same way that U arising from C deamination is removed. Cells that are dut^- ung^-, therefore, have a high level of uracil in their DNA (in fact, about 1 in 100 T is replaced by U).

These functions of the *dut* and *ung* genes can be exploited in mutagenesis (Figure 7.7). Template DNA is prepared from dut^- ung^- cells. It contains high levels of U. The template is used with the mutagenic oligonucleotide as described before (using dATP, dCTP, dGTP and dTTP) in a second-strand synthesis or a PCR reaction, and the products transformed back into *E. coli*. However, this time, the host strain is ung^+, and the activity of the glycosylase and the apyrimidinic-site endonuclease cause widespread degradation of the original DNA, as it contained U residues. Newly synthesized material (bearing the mutated sequence) is preserved, as it does not contain U residues. This method has the disadvantage that the dut^- ung^- hosts used for template preparation have an increased spontaneous mutation rate.

4. **Using methylation to target wild-type molecules**. In the previous example we used the *ung* system to direct the degradation of the starting template selectively. We can use a methylation-dependent system in a similar way. There are a number of variations of this system, but they rely on using methylated DNA as the template and then treating the products with enzymes that specifically degrade methylated or hemimethylated DNA. Methylated template DNA can be obtained by isolating it from a host strain that has methylase activity, e.g. from the Dam methylase (which methylates A residues). Alternatively, DNA can be methylated in vitro by treating it with methylase. After second-strand synthesis or PCR, DNA molecules that comprise two of the original template strands are fully methylated, i.e. methylated on both strands. DNA molecules containing one template strand are hemimethylated, i.e. methylated on that strand. Newly synthesized DNA is not methylated at all. The DNA is then treated in such a way as to damage methylated strands. This can be done by introducing it into a host strain that contains the enzymes Mcr and Mrr, which specifically degrade methylated DNA. Alternatively, it can be treated in vitro with a restriction enzyme

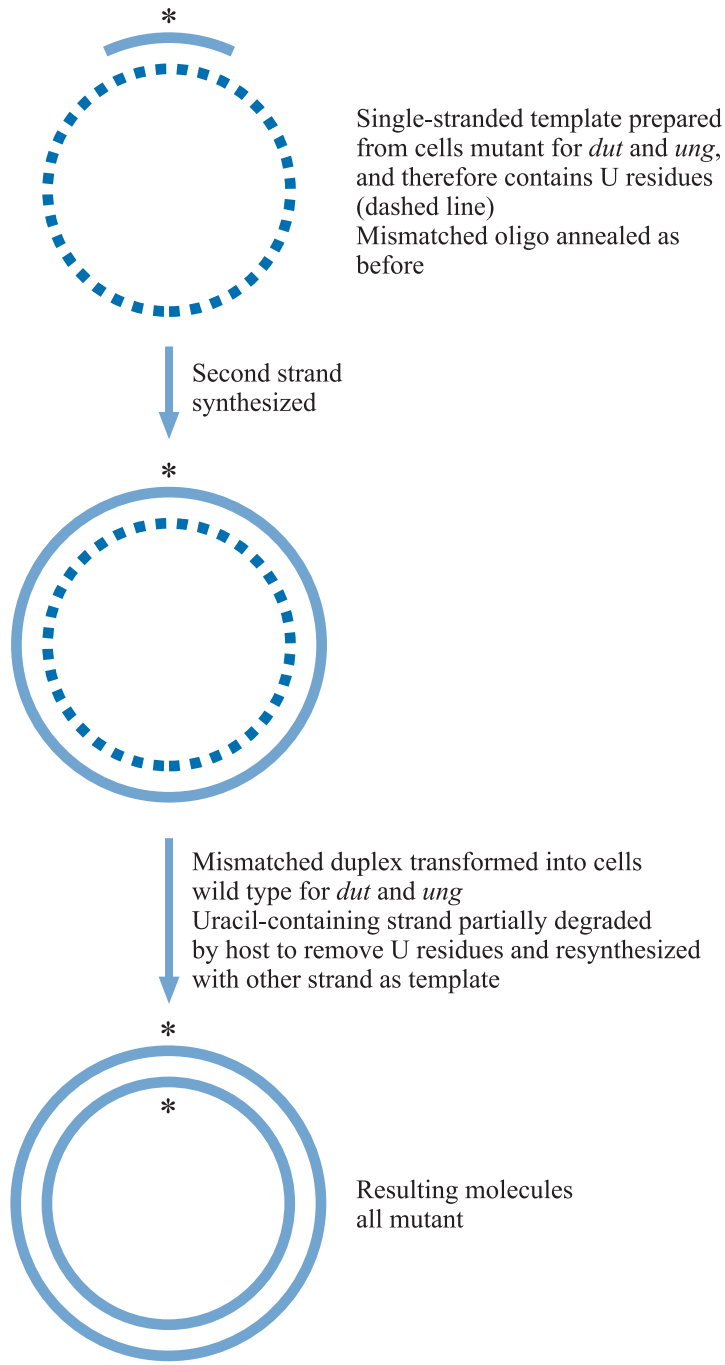


Fig 7.7 Mutagenesis using template prepared from *dut*$^-$ *ung*$^-$ cells (dashed circle). The asterisk indicates the mutation we need to make.

such as *Dpn*I, which specifically cuts methylated strands. The DNA is then introduced into *E. coli*. Whichever approach is used, the methylated and hemimethylated DNA is destroyed in the host, leaving the unmethylated DNA, which should carry the desired mutations.

7.4 Choosing the right mutations

The most important aspect of any site-directed mutagenesis project is not the generation of the mutants, which ought to be straightforward, but deciding which mutations to make. Careful consideration must be given not only to which nucleotide to mutate, but what it should be mutated to. To do this properly requires detailed knowledge of the function of the target sequence. If the sequence encodes a protein, then some idea of the protein's structure and of the residues that are important for function is usually necessary. Ideally, the three-dimensional structure of the molecule would be known. It is also important to recognize that minor changes in amino acid sequence can have a profound effect on the overall three-dimensional structure, perhaps even abolishing activity, even though there may be no change in the primary sequence at the active site. For example, even a small mutation might interfere with folding and lead to large-scale structural changes. Computer modelling may allow prediction of the effects of particular mutations on the three-dimensional structure. Once a mutant protein has been made, changes may also be detected by analysis of either nuclear magnetic resonance or circular dichroism, or by looking at antibody binding and protease resistance (which may be reduced in proteins that have not folded correctly).

7.5 Inactivating genes

It is often useful to inactivate endogenous genes in an organism. This might help us to find out the physiological role of the wild-type gene, or to direct the expression of a mutated gene in the absence of a background of expression of the wild-type gene. In a biotechnological setting, we might want to inactivate an undesirable gene. There are a number of ways in which this can most conveniently be achieved, involving either the disruption of the DNA sequence or the inactivation of the relevant RNA. The vectors, transformation systems and so on that are used depend on the organism involved (see Chapter 9), but the principles are largely the same and will be presented here. Note that before trying to inactivate a gene (whatever method is used) it is important to consider whether the result is likely to be lethal to the organism. If it is, the technique may superficially appear to have failed, since no gene-inactivated individuals will be recovered as they are all dead!

7.5.1 Databases of mutant lines

For many species, especially the popular model organisms, there are databases listing mutant strains. These strains may have been produced in systematic mutagenesis programmes or in individual experiments. Depending on how the database is organized, it may be

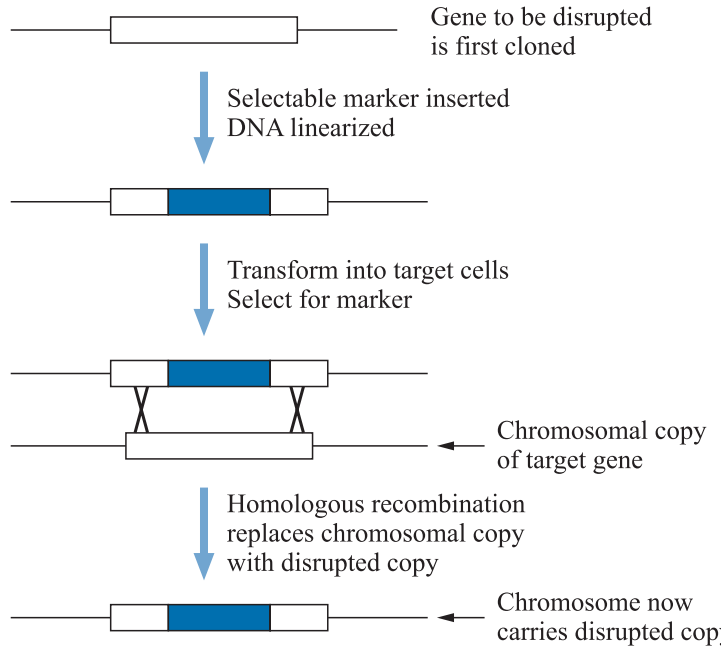


Fig 7.8 Gene disruption. Replacement with a disrupted copy.

possible to search it for strains with mutations in particular genes and order appropriate strains from a stock centre. It is advisable to check that the strain received *does* have the specified mutation (and no others), but this approach is usually much quicker than making the mutant line *de novo*.

7.5.2 Gene disruption

The principle of gene disruption is to use homologous recombination to replace the endogenous chromosomal copy of a gene with an inactivated copy. The procedure is shown in Figure 7.8. The gene to be disrupted is cloned and a selectable marker (perhaps an antibiotic resistance or a nutritional marker) is introduced into it. Insertion of the selectable marker must render the target gene non-functional, and it may be necessary to verify that this will indeed be the case. The disrupted gene is then excised from its vector (or at least the construct is linearized) and the linear DNA is introduced into the target organism. Selection is then imposed for the inserted antibiotic resistance or nutritional marker. The incoming molecule bearing the marker will not be stably replicated, as the molecule is linear. Stable acquisition of the marker can take place only if a double crossover across the flanking sequences and their chromosomal counterparts causes the marker's integration into the chromosome, replacing the endogenous chromosomal copy of the gene with the disrupted one. So, individuals that have stably acquired the marker should have the chromosomal copy of the target replaced by the disrupted copy.

It is not always necessary to supply the disrupting copy on a linear molecule. It can be supplied on a circular molecule, provided this is

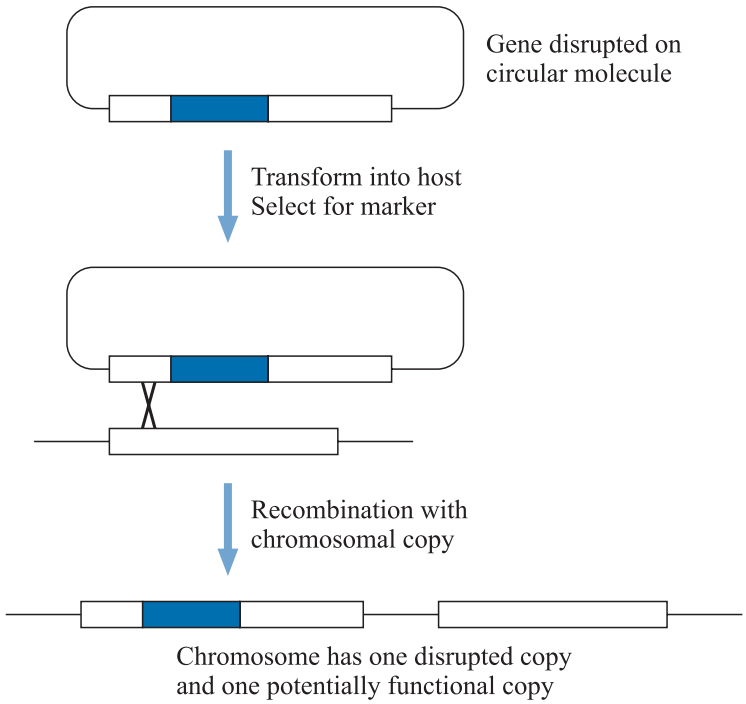


Fig 7.9 Gene disruption. Insertion of a disrupted copy may still leave a functional copy in the chromosome.

unable to replicate in the host used and can be stably maintained only by chromosomal integration. There are three possible outcomes.

(a) A double crossover may replace the chromosomal copy with the incoming disrupted copy as before.

(b) A single crossover results in a potentially functional copy (composed of part of the disrupting sequence and part of the chromosomal copy) adjacent to a disrupted copy, as shown in Figure 7.9. Integration in this way can be exploited to disrupt the endogenous copy of a gene and to insert an altered (but functional) one. Depending on the number of mutations in the incoming DNA, both copies may be non-functional.

(c) A single, non-homologous, crossover at a different site results in the integration of the incoming DNA with the endogenous copy remaining intact.

It is possible to select against outcomes (b) and (c) if necessary. Note that, in both of these, the entire incoming plasmid is integrated into the host chromosome. If this plasmid includes a marker whose *absence* can be selected (a counterselectable marker) then we can select against integration of the whole plasmid. For example, expression in *E. coli* of the *Bacillus subtilis sacB* gene, encoding levansucrase, is lethal in the presence of sucrose, as discussed in Chapter 4. If we are using an *E. coli* host, and the incoming plasmid contains an expressed *sacB* marker (in addition to the marker used for disruption of the target gene), then we can select against (b) and (c) simply by growth in the presence of sucrose.

Apart from the general problem of lethality discussed in the opening of Section 7.5, gene disruption has certain other limitations. If the target organism normally contains more than one copy of the gene, then it may be difficult to ensure that all copies are disrupted. This is especially true if the selectable marker is dominant (which is normally the case), since one copy of the marker will have the same effect as several. Multiple copies of the gene are likely to be present if the organism is not haploid or if there are several copies of the gene per haploid genome (i.e. the gene is part of a multigene family). An efficient homologous recombination system in the host is necessary to direct integration of the disrupted gene copy at the site of the endogenous gene, rather than elsewhere in the genome (although it may be possible to select against the latter, as described above).

Gene disruption is widely used in bacteria, yeast and animals. The particular problems posed by the need for homologous recombination in mammalian systems are discussed in Chapter 9.

7.5.3 Post-transcriptional gene silencing

A group of important methods relies on cellular processes that lead to inactivation of gene expression by affecting the RNA, i.e. post-transcriptionally. The processes involved are now reasonably well understood. The first method developed relied on antisense RNA, i.e. the modification of the host cell to synthesize an RNA molecule with a sequence complementary to the transcript of the gene to be inactivated. This is achieved by placing a DNA sequence encoding an RNA complementary to the RNA to be inactivated under the control of a suitable, powerful promoter and inserting it into the organism of interest. Expression from the promoter then generates antisense RNA and the level of expression of the endogenous gene is diminished. It is difficult to get complete inactivation this way, in contrast to the consequences of gene disruption, but levels of expression can be reduced to a few per cent or less of wild type. The technique does not suffer from the drawbacks for gene disruption (need for homologous recombination, problems of multiple gene copies, etc.) mentioned above. This approach has been particularly successfully used with plants (see Chapter 9). It was found that gene expression could also be reduced in some organisms by using shorter antisense oligonucleotides, rather than antisense RNA. In this case, the oligonucleotides are introduced directly into the target cells. The mechanism for suppression of gene expression using antisense RNA or antisense oligonucleotides was not clear originally. It was generally assumed that the mRNA and antisense RNA formed a hybrid that could not be translated, or was degraded by double-stranded-RNA-specific nucleases.

An unexpected observation with antisense RNA work with plants was co-suppression. This is the down-regulation of expression of endogenous genes by transformation with constructs that would generate *sense* RNA, rather than *antisense* RNA. At the time it was

suggested that the phenomenon was caused by the production of antisense RNA from the constructs, by readthrough from endogenous promoters located near to the site of insertion of DNA in the genome. Other causes, such as methylation of the endogenous gene, were also suggested. Around this time it was also shown that microinjection of RNA into *Caenorhabditis elegans* could suppress gene expression. Either sense or antisense RNA could be used, but unexpectedly large levels were required. It was then shown that the effect was due to a small amount of *double-stranded* RNA that was also present.

It is now recognized that, in a wide range of organisms, the presence of small double-stranded RNAs (where one strand has the same sequence as an endogenous RNA) leads to the breakdown of the corresponding endogenous single-stranded mRNA. In plants, and possibly in other organisms, this is thought to be a protection mechanism against viral attack, as many viruses produce a double-stranded RNA replication intermediate. The system may also be generally important as a part of a broader mechanism for protection of the genome against transposable elements, and for regulation of expression of endogenous genes by microRNAs. In general, the double-stranded RNA is cleaved by a nuclease called Dicer into short fragments of about 22 nucleotides, called short interfering RNAs, or siRNAs. These siRNAs are then recruited by a multisubunit complex called the RNA-induced silencing complex (RISC). This complex separates the two RNAs and uses one as a substrate guide to identify the corresponding mRNA, which it degrades. The short RNAs can also lead to silencing through other mechanisms, such as DNA methylation. Remarkably, in plants, the effect appears to be systemic. The presence of siRNAs to a virus in one part of a plant can lead to resistance of other parts of the plant.

The phenomenon of RNA inactivation by siRNAs is termed RNA interference (RNAi). It is now widely used in a broad range of organisms for the inactivation of genes by the introduction of short double-stranded RNAs from the gene in question. The siRNAs can be introduced directly, e.g. by microinjection or, in the case of *C. elegans*, by feeding. Alternatively, we can introduce into the host organism a DNA construct that, when transcribed, will generate a short RNA that is self-complementary. This RNA will fold back on itself to generate a short double-stranded species capable of activating the RNAi effect.

7.5.4 Cre-*lox* excision

The Cre recombinase protein of bacteriophage P1 mediates site-specific recombination at a 34 bp sequence, *loxP* (as discussed in Chapter 4). We can exploit this system to bring about controlled excision of particular sequences. One example of how this can be done is shown in Figure 7.10. The first step is to flank the target sequence with directly repeated *loxP* sequences. This can be done by replacing the endogenous copy of the target sequence with one flanked by *loxP* elements, which should not stop expression of the target sequence.

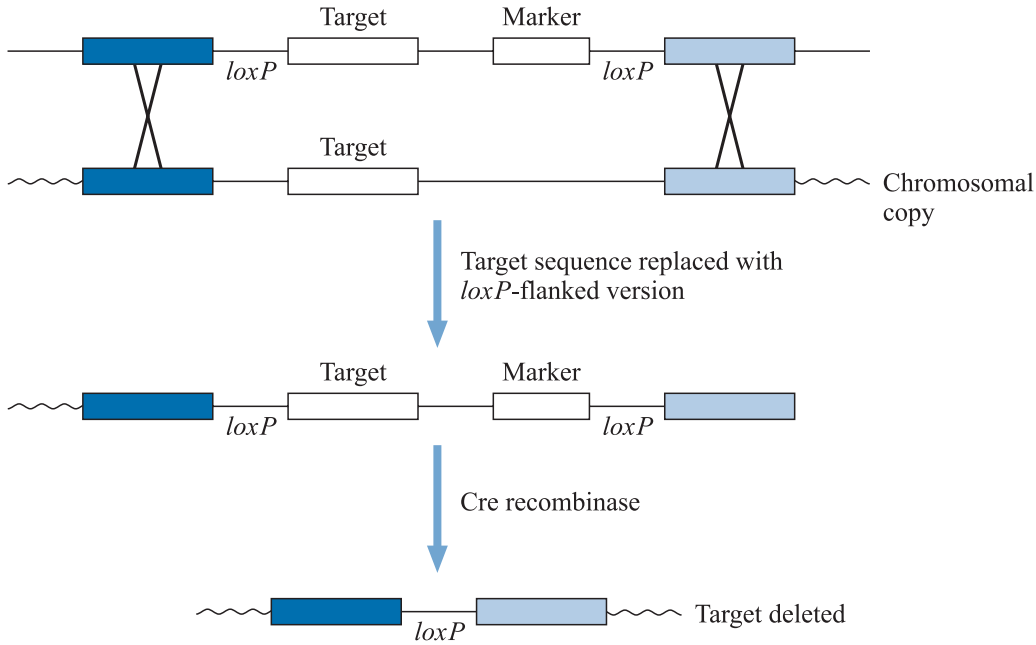


Fig 7.10 Cre-*lox* excision. The chromosomal copy of the target sequence is replaced with a copy flanked by *loxP* sequences. Supply of Cre leads to recombination across *loxP* and excision.

We then supply Cre recombinase. This could be achieved by separate integration of the *cre* gene under a controllable promoter followed by induction of the promoter. Induction results in expression of Cre, recombination across *loxP* sites and excision of the sequence between. We can also introduce *cre* by crossing with a strain containing the gene, or infection with a virus containing it. Notice that recombination across *loxP* as shown in Figure 7.10 resulted in the removal of the selectable markers used to set the system up. This is useful, in that it avoids potential disturbance of gene expression caused by having an actively transcribed selectable marker at the locus of interest. Although the *loxP* sequence is left, this is unlikely to have a major effect on the expression of surrounding sequence. Deletions generated in this way are sometimes referred to as subtle mutations. This system is also useful in removing selectable markers (such as antibiotic resistance genes), which may be the subject of public concern, from transgenic crop plants.

The ability to control the expression of Cre allows us to control when or where excision occurs. If we can direct time- or tissue-specific expression of Cre, then we will similarly get specific excision of the target. This may be very helpful if inactivation of the target throughout an organism is lethal. Excision of DNA flanked by *loxP* sequences is sometimes known as floxing. Other site-specific recombination systems can also be used, such as the yeast FLP system.

7.5.5 Ribozymes

In some circumstances, RNA molecules can have direct catalytic activity, such as the removal of certain introns from RNA by self-splicing. The term ribozyme is often used to describe any sort of RNA

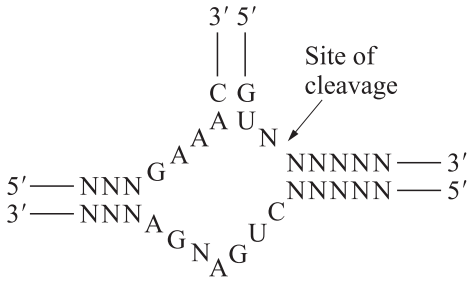


Fig 7.11 Ribozyme self-cleavage. Cleavage takes place within the 'hammerhead' structure shown. The sequence requirements for the cleaved molecule are minimal.

molecule that shows catalytic activity. However, the term is sometimes used more specifically to refer to RNA molecules that carry out a self-cleavage reaction, and these are the examples we consider here. Many of the examples come from RNAs that are associated with plant pathogenic viruses and viroids, such as the tobacco ringspot virus or the avocado sunblotch viroid. However, similar instances have also been reported from other systems, such as the hepatitis delta virus (which is usually associated with hepatitis B virus) and transcripts of some tandemly repeated satellite DNA sequences from animals. The reaction catalysed is an internal nucleophilic attack by a 2′-hydroxyl of the RNA on the phosphate group of the sugar–phosphate chain at the cleavage site. Because it involves the 2′-hydroxyl, this reaction is necessarily confined to RNA rather than DNA. In vivo this reaction is probably used to cleave into monomers molecules that are produced as concatemers by a rolling circle mechanism; the monomers are subsequently turned into circular monomers by ligation. The sequence requirements for the cleavage are relatively few, but they are always found within a region that is capable of forming a 'hammerhead' structure. Figure 7.11 illustrates this structure and also indicates the cleavage site. It is possible for the substrate molecule to be a separate one from the rest of the hammerhead, as long as the correct base-pairing can form. In other words, the sequence requirements for the substrate molecule are very small indeed. Therefore, if we know the sequence of a given RNA molecule, we should be able to tailor a catalytic molecule that can form a hammerhead with the first molecule and cleave it. A gene that will direct the synthesis of this catalytic molecule can, therefore, be constructed and introduced into a host organism, and any suitable substrate RNA molecule should be cleaved. It is also possible to supply the RNA exogenously, depending on the species under study. This technology is of particular medical interest as a way of inactivating intracellular pathogens, such as viruses, or modulating the expression of endogenous genes.

Chapter 8

Use of cloned DNA

8.1 Use as DNA

Cloned DNA can be used directly in too many different ways to describe here. They include sequencing, blotting (Southern, northern, southwestern, etc.), transcript mapping, footprinting and bandshift (gel retardation) assaying. Details of these can be found in general molecular and cell biology textbooks. Cloned DNA can also be used to build microarrays, as described in Chapter 1. We will concentrate first in this chapter on the use of cloned DNA sequences for expression, i.e. for directing the synthesis of RNA or protein. As we shall see, many cloning vectors have been developed specifically for this purpose. They are called expression vectors. We will then look at the use of cloned genes as tools for studying the function of other sequences.

8.2 Synthesis of RNA

8.2.1 Why synthesize RNA?

It may be necessary to produce RNA for a number of reasons. We might be studying RNA processing events, such as splicing or cleavage, in vitro and, therefore, need to produce RNA of a single type in the presence of as few contaminating proteins as possible. Or the aim might be to make a protein in a radiolabelled form by translation in vitro of an appropriate RNA in the presence of radioactive amino acids.

8.2.2 Vector systems

One approach to making RNA might be to use a vector with a powerful promoter to direct transcription in a bacterial cell in vivo and then isolate the RNA. However, because the purification of RNA from bacterial cells is difficult, and the isolation of individual RNA species even more so, the production of RNA is usually done by transcription from cloned DNA in vitro. The most widely used

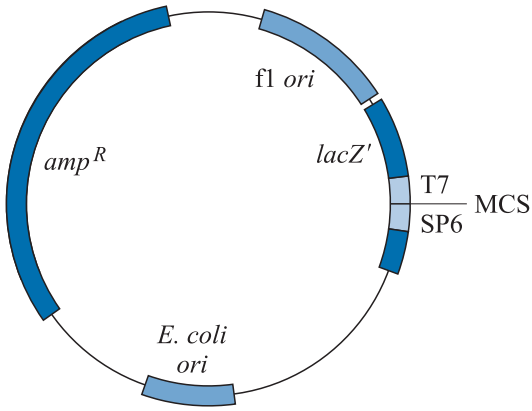


Fig 8.1 pGEMIIZf (3.2 kb). The vector contains a selectable ampicillin resistance gene (*amp*^{R}), origins of replication for double-stranded (*E. coli ori*) or single-stranded (f1 *ori*) DNA synthesis and a multiple cloning site (MCS) in a *lacZ′* gene. The MCS is flanked by T7 and SP6 promoters, allowing transcription in vitro of sequences inserted in either orientation.

systems utilize promoters from the bacteriophages T7 and SP6, but T3 promoters are also used. These phages encode novel RNA polymerases that are synthesized after infection of *E. coli*, and are very specific for phage promoters. A typical example of one of these vectors is illustrated in Figure 8.1. The vectors contain both T7 and SP6 promoters, orientated to direct transcription into the multiple cloning site that separates them. So it is always possible to transcribe a cloned sequence (regardless of the orientation in which it is cloned) by selecting the appropriate polymerase to activate the appropriate promoter. Transcription from the other promoter may be used to generate an antisense-strand RNA.

All that is needed for transcription is the incubation of the plasmid DNA in vitro with the appropriate polymerase and other substrates. Capping of the message can be achieved, if necessary, by incubation of the transcripts with the capping enzyme guanylyl transferase and GTP. If the first nucleotide to be incorporated into the transcript is a G, then it may be possible to bring about capping during transcription by the inclusion in the reaction of a cap analogue, G(5′)ppp(5′)G in excess over the normal nucleotide ppp(5′)G. The cap analogue can be incorporated only at the beginning of the transcript (because the 5′ groups, which would be needed for incorporation elsewhere, are in effect chemically blocked by each other), but it will be incorporated at the beginning of most transcripts because it is in excess over GTP. If it is important to avoid any transcription beyond the insert into the vector sequences, then the plasmid can be linearized by digestion with an appropriate restriction enzyme (cutting just beyond the site of insert integration) prior to transcription in vitro. Note that the example in Figure 8.1 has the site of insertion within the *lacZ′* section of the *lacZ* gene; therefore, as described earlier, we can verify the presence or absence of insert by the colony colour on plates containing X-gal. The plasmid also contains the filamentous phage f1 replication origin, in case we wish to generate single-stranded DNA. If cells containing the plasmid are subsequently infected with an appropriate helper phage, then

single-stranded DNA can be produced, packaged and secreted into the medium.

8.3 Synthesis of protein

8.3.1 Why synthesize protein?

It is extremely common to need to express cloned genes, so we can make the proteins they encode. We may need to prepare small quantities of radiolabelled proteins, perhaps for studies on co- or post-translational targeting or modification of proteins. Alternatively, we may want to produce larger amounts of the protein in a non-radioactive form, perhaps for determination of some of its properties in biological or biochemical assays, or for crystallization for structural studies. Or we may want to prepare an altered protein after site-directed mutagenesis to ascertain the consequences of the mutations we have generated for the catalytic properties of an enzyme.

8.3.2 The problems encountered

If we simply need to prepare small quantities of a radiolabelled protein, this is usually most easily accomplished by translation in vitro of RNA produced by transcription in vitro (see above). This is generally relatively straightforward. Occasionally, translation of small quantities of RNA is carried out in vivo, e.g. by microinjection into *Xenopus* oocytes. Larger quantities of unlabelled proteins are usually prepared by transcription and translation in vivo, and there are a number of problems to be overcome with these procedures. We need to consider the following:

(a) transcription and translation initiation
(b) fate of the transcript
(c) efficiency of translation
(d) fate of the protein after synthesis.

Methods for dealing with these considerations depend on the vectors used.

8.4 Translation in vitro

There are three commonly used methods of directing protein synthesis in vitro. They are based on a lysate of reticulocytes (immature red blood cells), a wheat-germ extract and an extract from *E. coli*. The reticulocyte lysate used comes from rabbits. Reticulocytes are prepared, lysed and the extract treated with micrococcal nuclease. This degrades the endogenous mRNA in the extract, which otherwise would produce a high background of translation products. The extract is then treated with

ethyleneglycol−O−O′−bis (2-amino-ethyl)−N, N, N′, N′−tetraacetic acid (EGTA), which chelates the calcium ions needed for the functioning of the micrococcal nuclease. If the nuclease is not inactivated, then the added mRNA will be degraded. Instead, added mRNA will be translated, and if a labelled amino acid (or acids) is included (usually radioactive methionine, although others, such as cysteine and leucine, are also frequently used) the protein produced will be labelled. The second commonly used system for directing protein synthesis in vitro is an extract of wheat germ. Basically, this is prepared by grinding wheat germ in a suitable buffer, removing cell debris, and treating with micrococcal nuclease followed by EGTA as with the reticulocyte lysate. With both these methods, transcription is usually carried out as a separate reaction (as described in section 8.2), prior to addition of the RNA to the reticulocyte or wheat-germ system. Some manufacturers produce systems in which transcription and translation systems are mixed, and the two processes are carried out together. This is also the case for the third method for translation in vitro, which uses an S-30 extract of *E. coli* cells. Cells of a suitable *E. coli* strain (usually deficient in RNase activity) are lysed, and the supernatant of a 30 000 *g* centrifuge spin is taken. This is the S-30 extract. It contains ribosomes and the other components of the protein synthesis machinery, but it does not contain the DNA, which was pelleted. Many of the ribosomes will be in the process of translating mRNA, so the extract is preincubated to allow the completion of translation that has already initiated. In prokaryotes, transcription and translation are usually coupled; that is, translation of the RNA is initiated while RNA is still being produced by transcription. Consequently, once translation of endogenous RNA has completed, it will not be reinitiated. Also, any fully transcribed RNA that is added to the system will not be translated efficiently. So, rather than transcribe RNA separately and then add it to the S-30 extract as we often do with the reticulocyte lysate or wheat-germ systems, we add DNA to the extract. Remember that the extract contains the machinery for transcription as well as translation. Any genes that have a promoter recognized by the S-30 extract (i.e. a typical *E. coli* one) can, therefore, be transcribed and translated. As with the other systems, a labelled amino acid is supplied and labelled protein is produced.

8.5 Expression in vivo

In this section we will concentrate on transcription and translation in *E. coli* cells. Similar principles apply when using other host species. We will not discuss other species in detail here, although some important considerations are detailed in the next chapter.

8.5.1 Requirements for initiation of transcription and translation

To direct initiation of transcription of a coding sequence, we need a promoter. However, it is often not enough simply to attach the coding sequence we want to express to a powerful, unregulated promoter, as many proteins are toxic to *E. coli* cells when they are over-expressed in this way. If the overexpressed protein is toxic, then no colonies will be recovered after transformation with the recombinant plasmid, because any transformed cells will be killed by the overexpression of the protein. It is possible that a few colonies might be recovered, with mutations in the cloned gene that either abolish the expression of the gene or modify the characteristics of the protein produced so that it is no longer toxic (but, in addition, no longer the protein that we are interested in). So it is usually preferable to use a powerful, but controllable, promoter.

As well as directing *transcription* initiation, we need to be able to direct *translation* initiation. In *E. coli*, this is done by a ribosome binding site. This is composed of a sequence of a few nucleotides in the mRNA, typically -GAGG-, shortly before the translation initiation codon. It is complementary to a few nucleotides at the 3′ end of the *E. coli* 16S rRNA. That sequence allows the small subunit of the ribosome to bind to the mRNA by complementary base-pairing. Translation starts at the first AUG codon in the mRNA downstream from the ribosome binding site.

So, expression of a cloned sequence in *E. coli* requires a promoter, a ribosome binding site and an initiation codon. There are a number of vectors, called expression vectors, designed to facilitate expression of cloned sequences. Examples of the strategies used are shown in Figure 8.2. Expression vectors usually contain a controllable promoter and may also provide a ribosome binding site and an initiation codon. If all three features are provided, then a hybrid protein is produced of which the N-terminal region is encoded by the vector and the rest is encoded by the sequence we inserted into the vector. The reading frame in the vector is in phase with the insert and runs directly into it. Translation will run directly from the coding region of the vector into the coding region of the insert. Proteins structured in this way are called fusion proteins. It is also possible to generate fusion proteins in which the C-terminal region is encoded by the vector. We will look at the promoters most commonly used, and then consider fusion proteins in more detail. Many expression vectors also carry powerful transcription terminators, such as that from the *E. coli rrnB* operon for rRNA, downstream from the cloning site. This is to stop high levels of transcription of other regions of the vector, such as the origin, which might interfere with the stability of the vector.

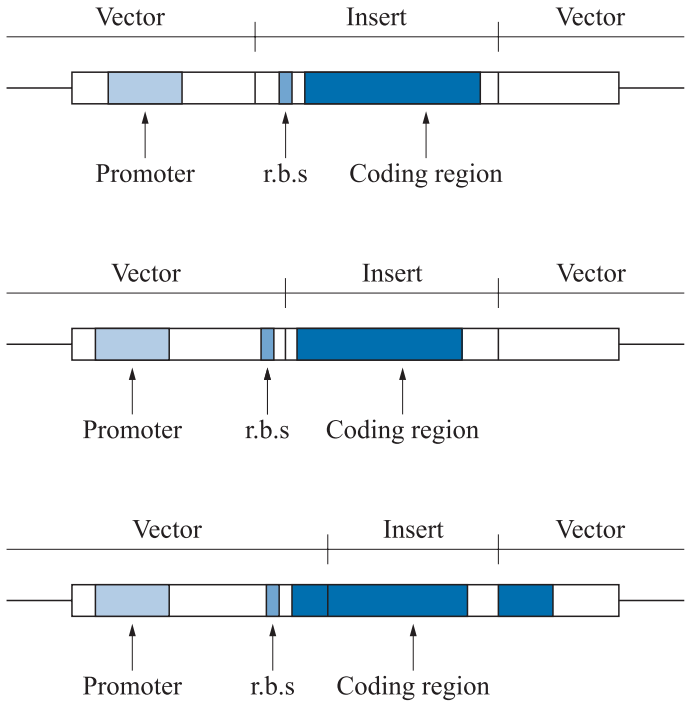


Fig 8.2 Expression of cloned sequences. r.b.s. = ribosome binding site.

8.5.2 Promoters

Several different promoters are available. Some of these occur naturally; others have been modified to improve their strength or the degree of control available. Many vectors use a two-stage control system, in which the activity of the promoter used for expression is modulated by a protein expressed from a different promoter, which is the promoter we directly control experimentally. Examples of these are given in the discussions of the lambda P_L promoter and the T7 promoter. These systems sometimes offer better control characteristics than those using a single promoter that is controlled directly. The promoters described here are available either as part of expression vectors or as cassettes that can be inserted into other vectors.

1. **lacZ promoter.** We have already come across one example of a powerful controllable promoter. This is the promoter of the *lacZ* gene, which we have already seen used in the pUC and M13 vectors, and the lambda ZAP and Bluescript series. In some vectors, a mutant form, called the *lacUV5* promoter, is used. This carries point mutations in the −10 region of the promoter, which increase its efficiency. The promoter is controlled by the LacI repressor. This is often encoded in the vector, or it can be provided by the host. In the presence of IPTG, the repressor function is inactivated and expression from the promoter occurs. Sequences inserted into the multiple cloning site in the *lacZ* gene will then be expressed. A difficulty may arise where the vector under

consideration is present in the cell at a high copy number and the repressor gene is present on an F′ plasmid, or the chromosome, at a much lower copy number. The multiple copies of the vector may be sufficient to titrate out all the repressor and allow expression in the absence of inducer. This problem of unwanted expression can be largely avoided by use of the *lacI*^{q} allele of the repressor. This carries a mutation in the promoter for the repressor itself, resulting in about 10-fold higher levels of repressor per cell than are obtained with the wild-type repressor allele. In many vectors the Lac repressor gene is carried on the vector molecule itself, which also helps to increase the amount of repressor in the cell and allows greater flexibility in the choice of host strain.

2. **Lambda P_L promoter.** The powerful leftward promoter (P_L) of bacteriophage lambda is tightly controlled by the lambda CI repressor protein acting through the O_L operator. The promoter directs transcription of phage genes during the infection process, but is widely used in expression vectors. The promoter can be controlled in a number of different ways. Perhaps the easiest way is to use a host that produces a mutant CI repressor protein (CI^{857}) that is temperature sensitive. At 30°C or lower, the repressor produced by the host is active and the promoter is kept under repression. Raising the temperature inactivates the repressor, and the promoter is activated. Another way of activating the promoter is to add nalidixic acid to the growth medium. This inhibits DNA gyrase and causes induction of the *E. coli* RecA protein, resulting in cleavage of the lambda repressor and consequent activation of the promoter. A third way of controlling the promoter is to regulate the production of the CI protein, by placing it under the control of another promoter that can be regulated. When the latter promoter is inactivated, production of CI protein stops, and expression from the lambda P_L promoter can begin. This is an example of the two-stage systems referred to in the introduction to this section.
3. **T7 promoters.** As described above, phage T7 encodes a polymerase that is very active and very specific for promoters from the same phage. A range of expression vectors has, therefore, been constructed containing phage T7 promoters to drive transcription of inserted sequences. The promoters for genes 5 and 9 of phage T7 are often used. Expression requires the production of phage T7 RNA polymerase. This is commonly supplied by a T7 RNA polymerase gene under the control of the *lacUV5* promoter. (In some systems it is under the control of a lambda P_L promoter.) Addition of IPTG activates the expression of the T7 RNA polymerase (when the *lacUV5* control system is used), and thus transcription of sequences under the control of the T7 promoter (Figure 8.3). This is another example of the two-stage systems referred to in the introduction to this section. Transcription of other genes in the host can be reduced by the addition of rifampicin, which inhibits *E. coli* RNA polymerase, but not T7 RNA polymerase.

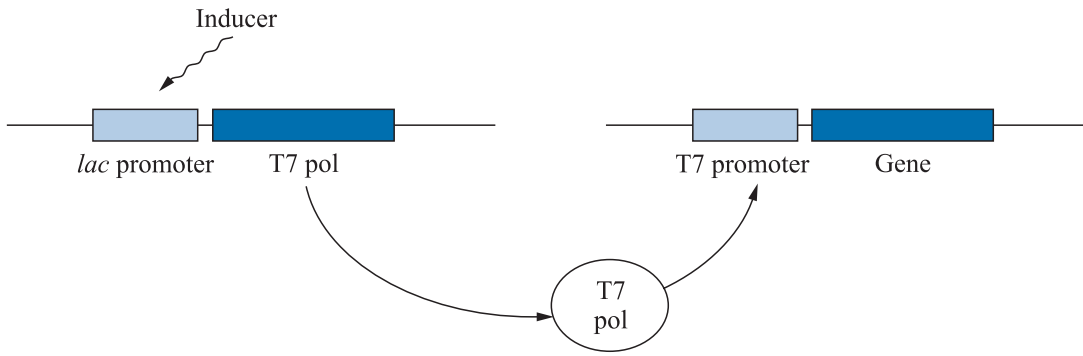


Fig 8.3 Two-stage expression systems. The target gene for expression is under the control of the T7 promoter. The T7 polymerase gene is under the control of the *lac* promoter and can be induced by IPTG. This leads to expression of the target gene.

4. **araBAD promoter.** The *araBAD* operon encodes proteins involved in arabinose metabolism. The promoter is controlled by the AraC transcriptional regulator and can be induced by addition of arabinose to the growth medium of cells containing vectors using this system. The *araC* gene is usually present on the vector.
5. **Hybrid promoters.** A number of powerful promoters have been constructed artificially as hybrids between existing promoters. One such is the *tac* promoter, which is a hybrid between the −35 region of the promoter of the *trp* operon (for tryptophan biosynthesis) and the −10 region of the *lacUV5* promoter. The *tac* promoter includes the *lac* operator and is regulated by the lac repressor, which has to be supplied by the host. Transcription is induced by addition of IPTG to the medium. Another hybrid is the *trc* promoter, which differs from the *tac* promoter in having the −10 and −35 sequences one nucleotide further apart. There is also a T7 promoter–lac operator hybrid, comprising a T7 promoter and the lac operator. Expression of this requires both T7 RNA polymerase and an inducer such as IPTG. Basal levels of expression, in the absence of the inducer, are very low indeed.

8.5.3 Fusion proteins

As explained above, expressing a sequence as an N-terminal fusion protein removes the need to provide a ribosome binding site and an initiation codon in the inserted sequence, as these are provided by the vector. Expressing a protein as a fusion may have other advantages. It may enhance stability, solubility, folding and disulphide bond formation. Antibodies to the vector-encoded part of the protein may be readily available and can be used to probe western blots and thus detect expression of the protein. Expression as a fusion to a periplasmic protein may allow the protein of interest to be targeted to the periplasmic space, which may be helpful for purification.

Affinity purification via the vector-encoded part of the protein may also be useful in purification. Several possibilities for the generation of fusion proteins are available. A short peptide region (usually of a few residues) fused to a protein of interest is often referred to as a tag.

1. **Glutathione-*S*-transferase (GST).** This enzyme catalyses the conjugation of glutathione to other molecules and has a protective function in many organisms. It is widely used as a basis for fusion proteins. The enzyme binds glutathione, allowing glutathione affinity purification matrices (such as glutathione attached to agarose) to be used very effectively in the purification of GST-fusion proteins.
2. **Maltose binding protein.** This protein is the product of the *malE* gene of *E. coli*. It is usually located in the periplasmic space, and functions in uptake of maltose and related sugars. Fusion proteins containing the leader sequence of MalE will usually be targeted to the periplasmic space, which aids in purification. Affinity purification can be carried out using a matrix containing amylose, a maltose polymer.
3. **Thioredoxin.** This protein contains two cysteine residues, which can be reversibly oxidized to cystine. The protein regulates the oxidation state of −SH groups on a range of other proteins. Fusion protein vectors are usually based on the *E. coli* gene *trxA*. Purification of fusion proteins is usually achieved by additional incorporation of a histidine tag (see below).
4. **Histidine tags.** In this example, the sequence to be expressed is fused to a sequence encoding a run of histidine residues (typically six) rather than to a naturally occurring protein. This can be either an N-terminal or a C-terminal fusion. Proteins containing histidine tags can be affinity purified using a matrix containing bound nickel ions. This is most commonly nitrilo-triacetic acid agarose (NTA-agarose). The bound nickel ions are chelated by the histidine residues of the tags (Figure 8.4).
5. **Pectate lyase.** This protein, PelB, is produced by some plant pathogenic bacteria and helps to degrade the plant cell wall. Fusion proteins produced in *E. coli* are usually secreted into the periplasm.
6. **Beta-galactosidase.** Fusion proteins incorporating beta-galactosidase can be purified using a matrix containing a covalently bound, non-cleavable, beta-galactosidase substrate analogue (such as *p*-aminophenyl-β-D-thiogalactosidyl-succinyldiamino-hexyl-Sepharose), or using a matrix containing β-galactosidase antibodies.
7. **Chitin-binding domain.** This is a 5 kDa domain from the C-terminus of a chitinase from *Bacillus circulans*. Fusion proteins carrying this domain can be purified using a column containing immobilized chitin.

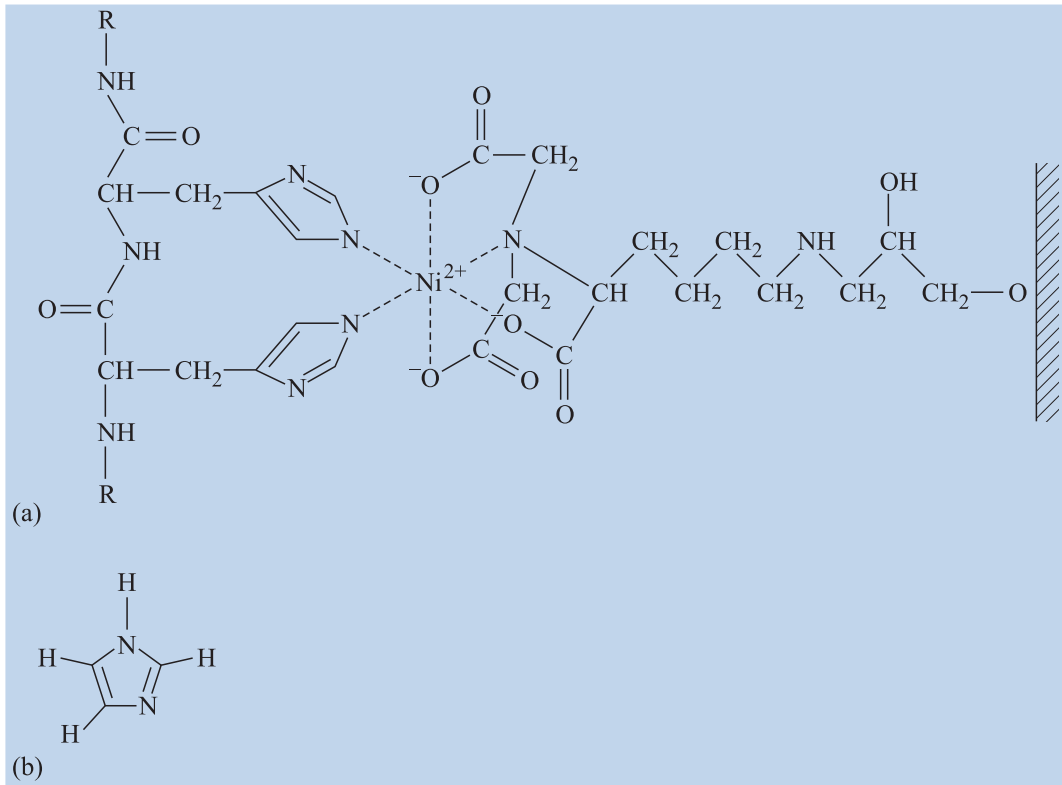


Fig 8.4 Purification of histidine-tagged proteins. (a) Histidine residues bound to nickel ions on an NTA column. (b) Imidazole, used to displace bound protein.

8. **FLAG tags.** The FLAG tag is an eight-residue peptide, DYKDDDDK. This sequence is very hydrophilic, so it is less likely to make the expressed protein insoluble. In addition, the last five residues are the recognition site for the protease enterokinase, which may be useful in removal of the tag after protein purification (see below).

Other fusion and tag systems include fusions to TrpE (involved in tryptophan biosynthesis), part of the Myc protein, the Nus protein of *E. coli*, the 'S' peptide derived from RNase A, and coloured proteins (which may facilitate monitoring of subsequent protein purification), such as haem-binding domains.

8.5.4 Purification of fusion proteins

Many fusion proteins can be conveniently purified from a cell extract by retention on a suitable affinity matrix. However, this leaves the problem of how to free the protein from the matrix. There are two common strategies for this. One strategy is simply to elute the fusion protein. This is usually done by washing the matrix with a buffer containing something that will compete for binding to the matrix or to the fusion protein. So, GST-fusion proteins can be eluted from glutathione-agarose with free glutathione, MBP-fusion proteins can be eluted from amylose matrices with free maltose, and his-tagged proteins can be eluted from NTA matrices with imidazole (see Figure 8.4), which competes with histidine for binding to the immobilized nickel. His-tagged proteins can also be eluted using

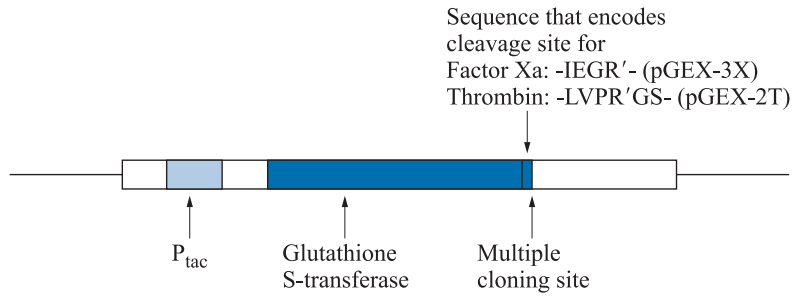


Fig 8.5 Expression regions of vectors that include protease recognition sites. The plasmids pGEX-3X and pGEX-2T (4.9 kb) are shown as examples. The vectors also have an origin of replication, an ampicillin resistance selectable marker and a *lacI*^{q} gene to supply repressor for control of the *tac* promoter (P_{tac}). The vectors differ in the cleavage site that is encoded adjacent to the multiple cloning site.

EDTA or a pH gradient to disrupt the chelation. The second strategy for releasing fusion proteins is to exploit specialized vectors that encode a protease recognition site at the junction between the fusion protein and the inserted sequence (as shown in Figure 8.5). The protease recognition sites commonly used are those for Factor Xa (-Ile-Glu/Asp-Gly-Arg-), thrombin (-Leu-Val-Pro-Arg-Gly-Ser-) or enterokinase (-Asp-Asp-Asp-Asp-Lys-). The first two proteases are involved in the blood clotting pathway; enterokinase is an intestinal enzyme that converts inactive trypsinogen to trypsin. After a fusion protein containing one of these sites has been adsorbed to a suitable matrix, the protein can then be released by treatment with the appropriate protease (Figure 8.6). It may be necessary to remove the protease subsequently, and problems can sometimes be caused by the protease cleaving the fusion protein at other sites as well. A third strategy for recovery of proteins from columns exploits the ability of certain proteins to carry out self-splicing. (This is a post-translational event, and is quite different from RNA splicing.) The splicing results in the excision of peptides called inteins and the ligation of flanking sequences (exteins). The strategy (Figure 8.7) uses an expression vector that generates a fusion protein with a modified intein sequence linking the protein of interest and the affinity tag. The intein has been modified so that it will self-cleave, on treatment with reagents (such as dithiothreitol) containing thiol groups, but will not bring about the ligation reaction. The fusion protein is purified on an appropriate affinity column and then treated to activate the intein self-cleavage. This liberates the remainder of the protein of interest.

There are also methods for chemical cleavage of the junctions within fusion proteins, e.g. with cyanogen bromide (which cleaves after methionine residues), but these are less widely used.

8.5.5 Co-expression of proteins

Sometimes it is desirable to express two or more proteins simultaneously. For example, they may form a complex that is difficult to reconstitute from the separate, independently isolated, proteins. If the proteins can be expressed in the same cell at the same time, they may assemble to form the complex, which can then be purified. There are two methods for this. One is to use separate

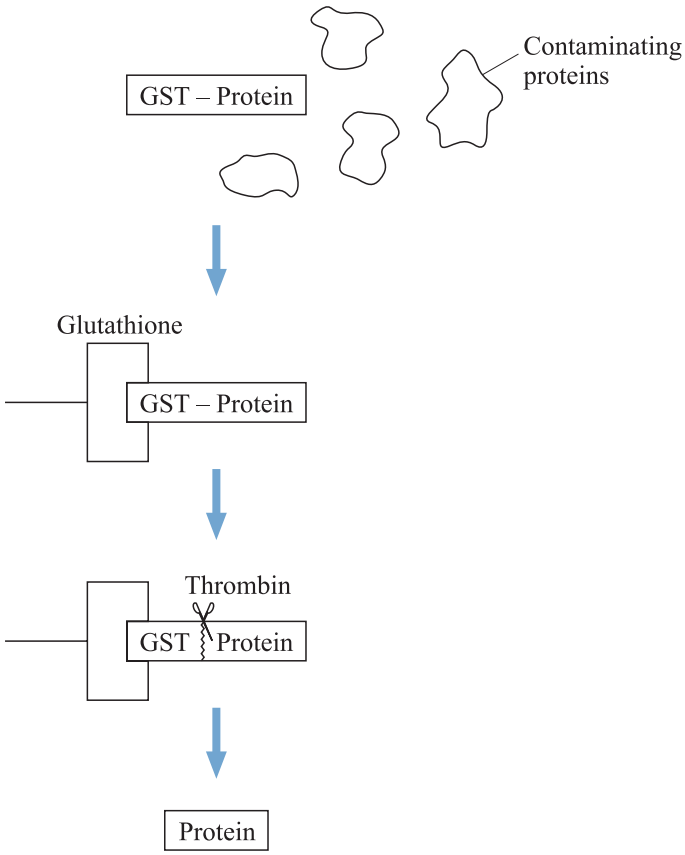


Fig 8.6 Purification and cleavage of a fusion protein. In this example, a GST fusion protein is purified by affinity chromatography on a glutathione matrix. Cleavage with an appropriate protease (thrombin in this example) releases the protein of interest.

expression vectors, with several in the cell at the same time. However, it is often difficult to maintain different, but related, plasmids in a cell stably. The second method is to use expression vectors that contain multiple independent sets of promoters and other signals in a single molecule.

8.5.6 Optimization

When expressing proteins in vivo, we usually wish to get as high a yield as possible. Very often, we also want the protein to be correctly folded and biologically active. These aims are not always compatible, so it may be necessary to find a trade-off between them. For example, we may find that if we express a protein at high levels it forms insoluble aggregates in the cell, with the protein incorrectly folded (if folded at all). A number of parameters can influence the outcome of attempts at expression, and we will look at the most important ones.

1. **Fate of the transcript.** A considerable amount of modification of the coding region of RNA molecules can take place in eukaryotes before translation. Most notable is the excision of introns, but other reactions can take place too, such as the

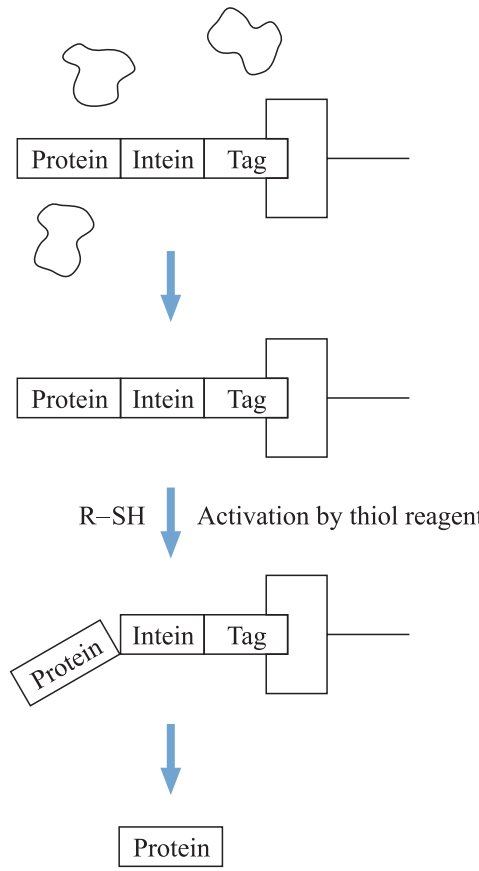


Fig 8.7 Intein-mediated cleavage of a fusion protein. The protein of interest and the affinity tag are linked by an intein domain. After affinity purification, self-cleavage of the intein is activated, liberating the protein of interest.

deletion, insertion or substitution of individual residues (editing). In general, these processes do not take place in *E. coli* (although introns are not completely absent from prokaryotes), so eukaryotic genes which normally require processing of their transcripts will not be expressed. The need for processing can be circumvented by expressing a cDNA molecule, rather than the gene itself. Assuming this cDNA has been generated from the fully processed message, there will be no need for further processing of transcripts generated from it.

2. **Efficiency of translation.** Assuming transcription and translation can be initiated efficiently, it cannot be assumed that the rest of translation will be efficient. In particular cases, strong secondary structure in a message may interfere with translation. Other factors that may influence the efficiency of translation are codon usage, codon meaning and the presence of introns (see above).

   Most amino acids can be encoded by a number of different codons, and not all codons are used equally frequently. Likewise, the tRNAs that recognize them vary in abundance. There may be differences in codon preference from one gene to another,

both within the same organism and between organisms. If the gene that is to be expressed contains many codons whose corresponding tRNA is present only at low abundance, then the overall rate of synthesis of the protein may be low, because the codons are read only with low efficiency. There are two solutions to this problem. One is to use site-directed mutagenesis to substitute the offending codons with codons that are read more efficiently. (In extreme cases, a whole gene can be synthesized to match the codon preferences of the host.) The second approach is to provide the host with a helper plasmid containing genes for the rare tRNAs. These helper plasmids are readily available commercially.

The genetic code is not universal, although deviations from it are relatively uncommon. However, if the gene we want to express comes from a system where the code is different, then expression in a system using the universal code is likely to lead to an altered product. This is perhaps most easily seen for those mitochondria (such as mammalian mitochondria) that use UGA as a tryptophan codon, rather than a stop codon. For example, expression of a mammalian mitochondrial gene containing UGA tryptophan codons in *E. coli* will lead to the formation of a truncated polypeptide, as the UGA codons are read as stop codons and protein synthesis is thereby terminated. To avoid this problem, site-directed mutagenesis can be used to change the UGA codons to UGG codons. Alternatively, one could make use of a 'chain-termination suppressor' strain of *E. coli*. Such strains contain mutated tRNA genes whose products have anticodons that are complementary to one of the termination codons. They can, therefore, 'read' a stop codon as something else, and in this particular example the UGA could be read as a tryptophan codon in an appropriate host. Because protein-coding regions frequently terminate with more than one different termination codon, the C-terminus of the protein should be unaffected.

3. **Toxicity of the protein**. Some proteins are toxic to cells when expressed at high levels. If this is so for the protein we are trying to express, then there will be a strong selective pressure for mutations in the host that inactivate expression or alter the protein itself. Therefore, it is useful to be able to keep basal levels of expression of cloned sequences as low as possible until we are ready to activate the production of our target protein. This will reduce the time available for the acquisition of undesirable mutations. Although the promoters used in expression vectors are controllable in principle, there are several addition precautions that can be taken.

   Promoters from genes involved in carbohydrate metabolism in *E. coli* are usually repressed in the presence of glucose. This is a natural physiological response to save a cell the biochemical cost of synthesizing enzymes to metabolize other sugars, when glucose is available. (The enzymes for glucose metabolism are synthesized

all the time.) So, if we are using an expression vector with the *lacZ* (or *lacZ*—T7 hybrid) or *araBAD* control systems, addition of glucose to the growth medium will help to repress basal transcription levels. When we are ready to commence expression, the cells are transferred to a medium lacking glucose.

Expression levels from T7 promoters in the absence of the T7 RNA polymerase are very low. When using a T7-based system, therefore, it may be helpful to propagate constructs in a host that lacks the T7 RNA polymerase gene. Only when we are ready to activate transcription are the constructs transferred to a host that can provide the T7 RNA polymerase. An additional way of reducing basal expression levels from T7 promoters exploits the T7 lysozyme. This phage-encoded protein has two functions. It attacks the peptidoglycan in bacterial cell walls and it also acts as an inhibitor of T7 RNA polymerase. Many of the hosts used for T7 expression vectors contain, in addition to the T7 RNA polymerase gene, a T7 lysozyme gene on a separate plasmid. The two most commonly used plasmids are called pLysE and pLysS. The former generates higher levels of the lysozyme protein in the cell than the latter. The lysozyme synthesized from these plasmids reduces the basal activity of the T7 RNA polymerase and, therefore, reduces the amount of transcription prior to induction. When synthesis of the RNA polymerase is induced, the lysozyme levels are no longer sufficient to inhibit the polymerase. A disadvantage of this system is that the reduction of T7 polymerase activity by the lysoszyme can mean that the rate of synthesis after induction is also reduced.

4. **Solubility and folding of the protein.** Many proteins expressed in *E. coli* do not remain soluble in the cell. Instead, they may form insoluble aggregates, termed inclusion bodies. This is not always a bad thing. The presence of the protein as inclusion bodies may facilitate purification, as the inclusion bodies may be easy to separate by centrifugation from the rest of the cell contents. The occurrence as inclusion bodies may also help to protect the protein against degradation, discussed in more detail below. It is possible to solubilize some proteins from purified inclusion bodies and then allow them to fold correctly. This is often achieved by solubilization using denaturing agents, followed by removal of the agents by dialysis or dilution into a buffer lacking them. The factors governing the formation of inclusion bodies are not clear. Their formation is generally favoured by higher growth temperatures and higher rates of synthesis. The protein to which the target protein is fused may also affect inclusion body formation, although if a foreign protein is synthesized as a fusion with an endogenous *E. coli* protein, then it is usually less likely to form inclusion bodies. Even if an expressed protein remains soluble, it is not guaranteed to fold correctly, or form correct disulphide bonds. If disulphide bond formation is important, then it may be

useful to direct the export of the protein to the periplasmic space. This is a more oxidizing environment than the cytosol, and thus favours disulphide bond formation.

5. **Stability.** Foreign proteins expressed in *E. coli* are often degraded. Why this degradation should occur is not always clear. It may be that if the protein concerned normally forms a complex with others, then it is unable to form a correctly folded structure in their absence. This results in partial denaturation and protease susceptibility. Formation of inclusion bodies or expression as a fusion protein often helps to maintain stability.

   A number of proteases are recognized as particularly important for the degradation of foreign proteins within *E. coli* cells. One of the major ones is the Lon protease. This protease is apparently particularly active against partially or completely denatured proteins. Inactivation of the *lon* gene can, therefore, be used to reduce the degradation of foreign proteins. This causes further difficulties, though, as the *lon* mutation can confer undesirable properties on the host cells. These include mucoidy and UV sensitivity, and sometimes reduced plasmid stability and transformation efficiency. Mucoidy is caused by overproduction of the polysaccharide capsule of the cells, making them difficult to manipulate physically. However, mucoidy can be suppressed by an additional mutation. This is often a mutation in the *galE* gene, or in the *cpsA–E* gene cluster (all of which are needed for synthesis of capsule polysaccharide). The tendency to mucoidy can also be reduced by alteration of the growth conditions, such as use of rich media and growth temperatures above 37°C. Other proteases that can lead to degradation of foreign proteins are the products of the *clpA* gene (a cytosolic protease) and the *ompT* gene (a protease in the outer membrane). Mutations in these are particularly useful in a background that is already lon^-. A number of proteases are under the control of the *htpR* heat-shock response gene, so mutations in this may also be helpful.

6. **Destination.** As already discussed, secretion of a protein from the cell into the periplasmic space (or even the growth medium) after synthesis may help to protect against proteolysis and will facilitate subsequent purification. Secretion may also aid in the formation of disulphide bridges. The formation of such bridges may not necessarily be an advantage, of course, depending on whether they are found in the protein in its functional state. Secretion generally (but not always) requires the presence of an N-terminal leader sequence on a protein, which is subsequently removed by a leader peptidase. In addition, for some proteins, sequences within the mature protein itself may have important effects on the efficiency of secretion. To direct the secretion of a foreign protein, therefore, we usually arrange for it to be synthesized as a fusion with the leader sequence of a secreted protein, such as MalE or PelB. Other *E. coli* proteins whose

leader sequences have been used in this way include beta-lactamase, the outer membrane proteins OmpA and OmpF, PhoA (alkaline phosphatase) and the filamentous phage gene III protein. Protein can also be exported from a cell as part of a fully formed filamentous phage, as in the phage display vectors described in Chapter 4.

7. **Modification.** Many proteins require modifications for biological activity, such as glycosylation, or the addition of cofactors, such as haem. Although prokaryotic expression systems will be able to make some of these modifications, they are unlikely to be able to carry out those that are restricted to eukaryotes. It is then necessary to use one of the eukaryotic systems discussed in Chapter 9.
8. **Other factors.** There are other factors that may be important in maximizing expression from a particular gene, particularly in industrial contexts. They include vector copy number and stability. Many of these factors are less well understood and, for most routine laboratory work using standard vectors, they are secondary considerations.

8.6 Studying gene function: reporter genes and tags

We saw in Chapter 6 how we can use reporter genes, whose activity we can readily measure, to screen libraries of cloned sequences for features such as promoters and terminators. We can also use reporter genes to study many aspects of gene function in vivo. The reporter genes most commonly used for these experiments are the fluorescent proteins (usually green, yellow, red or cyan), the luciferases, beta-galactosidase and beta-glucuronidase. The activity of all of these can be readily visualized and the activity of the different fluorescent proteins can be measured independently in the same sample. The gene for chloramphenicol acetyl transferase can also be used as a reporter, although it requires a biochemical assay, which is less convenient than direct visualization. We can also use the fusion sequences we looked at in this chapter as reporters. There are many ways reporter genes can be used. We will look at the general principles. Figure 8.8 summarizes the constructs we might make for these experiments. Many of these rely on the generation of transgenic organisms, which is discussed in Chapter 9.

8.6.1 Determining the tissue specificity of a promoter

We may want to know in which tissues of an organism the promoter of a gene of interest is active. We make a construct that places a reporter sequence under the control of that promoter and introduce the construct back into the host organism. The construct could be made either:

(a) by complete removal of the coding sequence and replacement by the reporter coding sequence, or

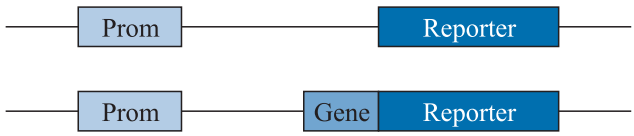


Fig 8.8 Reporter gene constructs. Both of these constructs could be used to determine promoter activity. The lower construct contains a fusion between the target gene and the reporter, and it could also be used to determine the subcellular location of the product of the target gene.

(b) so it encoded a fusion between the reporter sequence and the N-terminus of the protein whose gene we were interested in.

8.6.2 Determining the activity of altered forms of a promoter

We may have cloned a gene of interest and its associated promoter, and want to identify which regions of the promoter are important for correct expression. We make the desired changes in the promoter (Chapter 7), ligate the reporter gene to the modified promoter (either as a fusion or complete replacement, as above) and introduce the construct into the host organism. We then compare the level of expression of the reporter gene with what is obtained from an unmodified promoter.

8.6.3 Determining where a protein is located within a cell

We may want to know where a particular protein is targeted within the cell (perhaps to the mitochondrion or the chloroplast). We make a construct that contains a reporter coding sequence fused to the coding sequence of the protein of interest. Assuming (as is usually the case) the reporter sequence does not affect the targeting process, the fusion protein will be directed to its usual compartment, and we can determine this visually. It is not always easy to determine a subcellular compartment reliably when we are using fluorescence microscopy, but the ability to detect more than one kind of fluorescent protein in cells can solve this problem. We can use the green fluorescent protein as a reporter and express at the same time a red fluorescent protein whose subcellular location we know. We monitor the location of each separately. If the image of the green fluorescent protein can be superimposed accurately on the image of the red fluorescent protein, then the two proteins must be in the same place.

8.6.4 Gene trapping

We may want to identify sequences that are expressed in a particular tissue (or cell) type. A large number of transgenic organisms are created that contain a reporter gene integrated into the genome at different, random, locations. We screen the transgenic individuals for those that show the reporter gene expressed in the tissue of interest. These individuals, therefore, contain the reporter gene integrated into a sequence that is expressed in that tissue type. We can clone the sequences flanking the reporter gene using the transposon tagging approach described in Chapter 6. The use of reporter genes to identify expressed sequences in this way

is called trapping. A range of constructs is available to allow detection of enhancers, promoters and transcribed sequences, for example. These systems are then referred to as enhancer trapping, promoter trapping, and gene trapping respectively. Two examples are given in Figure 8.9. The first shows an enhancer trap system. The enhancer trap vector will not replicate independently in the host, but includes a selectable marker to allow us to select cells in which the vector has integrated into the host genome. The vector also contains a weak promoter, with a reporter gene adjacent to it. When the vector integrates near an enhancer, expression of the reporter is activated in those tissues where the enhancer is active. The second example in

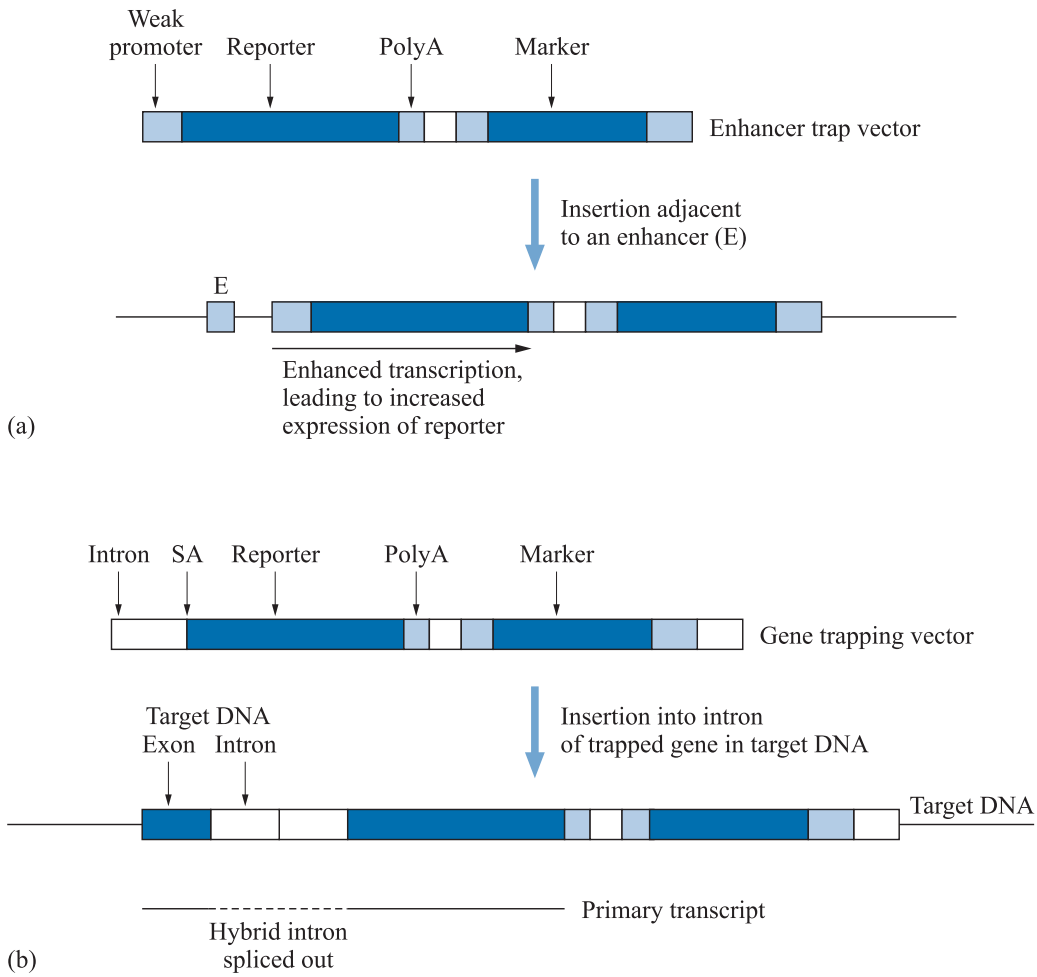


Fig 8.9 Trapping vectors. (a) An enhancer trap vector. The marker allows selection for integration of the vector. In the cases where the vector has integrated near the enhancer (E), there will be an enhanced expression of the reporter gene on the vector. (b) A gene trapping vector. The marker again allows selection for integration of the vector. In the cases where the vector has integrated within an intron (as shown), a transcript will be produced that contains a hybrid intron (dotted line), with part coming from the target DNA sequence and part from the intron in the vector. The hybrid intron will be spliced out using the splice acceptor site in the vector (SA), giving a fusion between the target DNA exon and the reporter sequence of the vector.

Figure 8.9 is a gene trap vector. Again, there is a selectable marker and a reporter gene. In this case, the reporter is promoterless and is preceded by an intron sequence. Insertion of the vector into an intron will give rise to a transcript that contains a hybrid intron, part coming from the target DNA sequence and part from the intron in the vector. The hybrid intron is spliced from the transcript using the splice acceptor site in the vector, fusing the reporter sequence to the target DNA exon.

Chapter 9

Using other organisms

9.1 Introduction

So far, we have concentrated on the techniques for manipulating the genome of *E. coli*. However, it is very likely that we will need to work with other organisms, too. We might be interested in some aspect of the biology of another species and want to study the effect of modifying it in some way. We might want to make genetically altered versions of commercially important species, such as crop plants, to improve their value. Or we might want to produce a protein in a form that requires some post-translational modification that *E. coli* is unable to accomplish. The principles that we have seen for *E. coli* apply in exactly the same way. We need to have suitable vectors, a means of getting the DNA into the organism and ways to select transformants. We may also need to take steps to increase expression. Often, we first clone DNA from one organism in *E. coli* and identify a recombinant containing a particular gene of interest. We then transfer that gene into some other host species to alter the host's properties. An organism containing a gene derived from elsewhere is said to be transgenic.

In this chapter we will look at the vectors and transformation systems available for a range of other organisms. We will look first at bacteria, and then we will consider fungi, plants (including algae) and animals. Rather than giving a detailed listing of all the methodologies for all the organisms (details of which can be found in primary research papers, reviews and laboratory manuals), we will concentrate on a representative sample of organisms to look at how the principles we have already encountered can be applied.

9.2 Bacteria

E. coli, which we have looked at in detail, belongs to the Gram-negative group of bacteria. This is just one of many bacterial

groups, all of which contain members that are important for clinical, economic, ecological or other biological reasons.

9.2.1 Vectors

The DNA used for manipulation of the species of interest could be in any of the following categories:

1. **Non-replicating DNA.** This can be linear material or circular material which, although perhaps able to replicate in *E. coli*, is not able to replicate in other hosts. It could, therefore, be stably maintained in the target species only by integration into another replicon such as the bacterial chromosome or another plasmid. If the host species has an active homologous recombination system and the incoming DNA contains sequence homologous to a site in the host genome, then insertion of the incoming DNA may take place at that site. Otherwise insertion may be at random sites, which may be disadvantageous.
2. **A plasmid unique to the species under study.** *E. coli* is, of course, not the only bacterial species to contain plasmids, and our target species may have plasmids of its own that we could use for genetic manipulation. However, it is more convenient if the plasmid used can replicate in *E. coli* as well as the target species. This would allow us to construct recombinant molecules using the well-developed systems available in *E. coli*, prior to introducing the molecules into the target species. This makes the next two categories more useful.
3. **A broad host-range plasmid.** There are many broad host-range plasmids that are able to replicate in a number of different bacterial species. They can be classified into incompatibility groups on the basis that the presence in a cell of a plasmid of one group inhibits the replication of different plasmids from the same group, as a result of interference between the processes controlling replication. There is a large number of incompatibility groups, with the five major ones being designated C, N, P-1, Q, and W. Members of the last three groups have been most widely exploited for cloning purposes, and examples are given in Table 9.1. Many have been modified by the incorporation of features such as multiple cloning sites and the *lacZ′* minigene. The plasmids differ in how easily they can be transmitted to other bacteria. Some are self-transmissible, in that they can direct their own transfer from one bacterium to another. Some of the others are unable to direct their own transfer, but can be transferred if functions are provided by another plasmid. These plasmids are said to be mobilizable. The rest cannot even be mobilized by other plasmids, and are said to be non-mobilizable. A plasmid that is transmissible may be easier to work with for that reason, but the plasmid is more likely to spread to wild populations in the event of release and, therefore, may be less desirable on safety grounds.

Table 9.1. Broad host-range plasmids of the P-1, Q, and W incompatibility groups

Plasmid	Group	Resistances[a]	Size (kb)	Transmissible?
RK2	P-1	Amp, Kan, Tet	56.4	Self
RP301	P-1	Amp, Tet	54.7	Self
pRK2501	P-1	Kan, Tet	11.1	Non-mobilizable (derivative of RK2)
pRK404	P-1	Tet (also *lacZ'*)	11.2	Mobilizable (derivative of RK2)
RSF101	Q	Str, Sul	8.9	Mobilizable
pKT254	Q	Str, Kan	12.8	Non-mobilizable
pGV1106	W	Sul, Kan	8.7	Mobilizable
pSa4	W	Cam, Str, Kan	9.4	Non-mobilizable

[a] Amp = ampicillin resistance, Cam = chloramphenicol resistance, Kan = kanamycin resistance, Tet = tetracycline resistance, Str = streptomycin resistance, Sul = sulphonamide resistance.

Table 9.2. Examples of plasmids from bacterial species and shuttle vectors derived from them

Plasmid	Source	Markers[a]	Size (kb)
pC194	*Staphylococcus aureus*	Cam	2.9
pHV33	pC194, pBR322	Cam, Amp, Tet	7.3
pIJ101	*Streptomyces lividans*	Colony morphology	8.8
pANT1200	pIJ101, pUC backbone (and other fragments)	Amp, Tsr	8.1
pUH24	*Anacystis nidulans* (cyanobacterium)		7.8
pSG111	pUH24, pBR328	Cam, Amp	12.9

[a] Amp = ampicillin resistance, Cam = chloramphenicol resistance, Tet = tetracycline resistance, Tsr = thiostrepton resistance.

4. **Shuttle vectors.** These are artificially constructed hybrids between *E. coli* plasmids and plasmids from other species. They are, therefore, able to replicate and be selected in both species. They are called shuttle vectors, as they can be shuttled from one host to another, and they are very widely used. Examples of some shuttle vectors are given in Table 9.2. In effect, they are plasmids whose host range has been artificially broadened by the introduction of an extra replication origin. Some shuttle vectors can replicate in several species. For example, the vectors based on the plasmid pC194 from *Staphylococcus aureus* can also replicate in *Bacillus* species. Typically, the initial construction and verification of a recombinant plasmid using a shuttle vector would be done using *E. coli* as a host, followed by transfer into the species of interest. There are shuttle versions of other kinds of vector, in addition to simple plasmids. These include vectors with phage elements, such as cosmids, and expression vectors. There are many markers in addition to the common ones we have already encountered. Some examples are given in

Table 9.3.	Examples of selectable markers used in prokaryotes
tsr	Resistance to thiostrepton (from *Streptomyces*)
hyg	Resistance to hygromcyin (from *Streptomyces*)
mel	Production of melanin (from *Streptomyces*)
trp	Tryptophan biosynthesis (from *Bacillus*)
sacB	Sensitivity to sucrose (from *Bacillus*)
neo	Resistance to neomycin, kanamycin (from the transposonTn5)
ble	Resistance to bleomycin (from the transposonTn5)

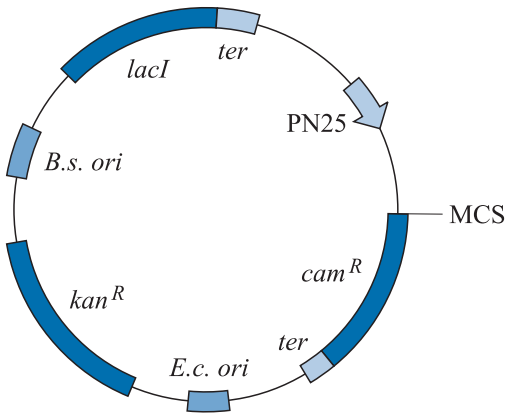


Fig 9.1 Plasmid pREP9 (7.3 kb), an example of a *Bacillus subtilis* expression vector. Sequences inserted into the multiple cloning site (MCS) are expressed from the *lac* operator–PN25 promoter (PN25). The plasmid can function as a shuttle vector, with origins for replication in *E. coli* (*E.c.*) or *B. subtilis* (*B.s.*). There are selectable markers for resistance to kanamycin (kan^R) and chloramphenicol (cam^R), and transcription terminators (*ter*).

Table 9.3. Some markers can be expressed effectively in a range of species; others may require the supply of expression signals from the species under study.

Expression vectors follow the same principles used for expression in *E. coli*. Many *E. coli* promoters will work in other Gram-negative bacteria, although powerful promoters from the species in question can also be used. Some promoters that work in *E. coli* will also work in more distantly related bacteria. For example, the P_{N25} promoter from the *E. coli* bacteriophage T5 will function in *Bacillus subtilis*, even though the latter is a Gram-positive species and *E. coli* is a Gram-negative species. If the P_{N25} promoter is fused to the *E. coli lac* operator, then the activity of the promoter can be controlled by the lac repressor in response to the addition of IPTG. The repressor is supplied by a *lacI* gene derived from *E. coli* but modified by attachment of a *B. subtilis* promoter and ribosome binding site to allow its expression in *B. subtilis*. An example of such a system, pREP9, is shown in Figure 9.1. It has a ribosome binding site and a multiple cloning site for DNA insertion, a chloramphenicol acetyltransferase reporter gene, a transcription terminator, a kanamycin resistance selectable marker, and Gram-negative and Gram-positive origins of replication (allowing use as a shuttle vector). A potential problem with expression in *B. subtilis* is

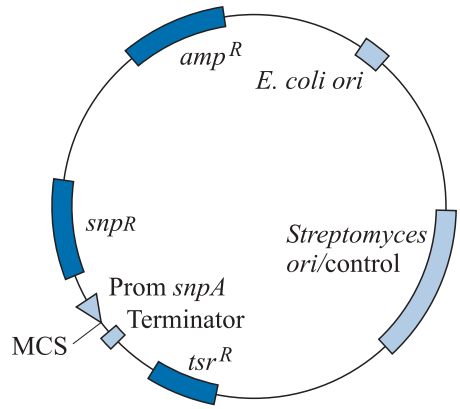


Fig 9.2 *Streptomyces* expression vector pANTI200 (8.0 kb). In addition to the ampicillin resistance gene (*amp*^{R}) and origin for selection and propagation in *E. coli*, there are an origin and an associated control sequence from *Streptomyces*, a thiostrepton resistance gene (*tsr*^{R}) and the small neutral protease promoter (Prom *snpA*) under the control of the product of *snpR*. There is a multiple cloning site (MCS) followed by a transcription terminator from the methylenomycin resistance gene.

that, during different growth stages (e.g. during sporulation), different promoters become active. Under these circumstances the expression of the promoters in a given expression vector may be considerably reduced. Homologous promoters can also be used, for expression purposes, e.g. from the *Bacillus* phage SPO-1. In *Streptomyces*, popular expression signals include promoters for *snpA* (small neutral protease), as shown in Figure 9.2, *aph* (aminoglycoside phosphotransferase), *ermE* (erythromycin resistance), *aml* (alpha-amylase) and *tipA* (for a thiostrepton-inducible protein).

Gram-positive bacteria are particularly attractive as hosts for secretion of expressed proteins into the growth medium, as they are able to do so at high rates. Secretion is usually achieved, as with *E. coli*, by expression as a fusion protein with the signal (leader) sequence of a secreted protein. Signal sequences used for *B. subtilis* include those for alpha-amylase, neutral protease, alkaline protease (subtilisin) and levansucrase. Although several of these come from different species (e.g. alpha-amylase and alkaline protease are produced by *Bacillus amyloliquefaciens*), the amino-termini of these proteins function satisfactorily in secretion by *B. subtilis*. A disadvantage with the use of *Bacillus* species as hosts for secretion is their production of high levels of endogenous proteases. Although hosts with reduced levels of protease are available, these hosts are not always satisfactory. Endogenous proteases may be less of a problem with *Streptomyces* species. Secretion signals used with these include those from tendamistat (an inhibitor of alpha-amylase, secreted by *Streptomyces tendae*) and others from proteases and protease inhibitor proteins.

9.2.2 Introduction of DNA

There are three main ways by which DNA can be introduced into recipient bacterial cells. These are the direct introduction of naked DNA into intact cells, transformation of protoplasts and

conjugation. In a few cases, and with appropriate vector systems, phage delivery may also be possible. Different methods are suitable for different species. It is important to bear in mind that many bacterial species contain restriction systems that may degrade incoming DNA. For example, these include the LT, SA and SB systems of *Salmonella*, which are analogous to the Class I *hsd* system of *E. coli*. There are Class II systems in a very wide range of bacteria, as witnessed by the very large number of different restriction enzymes available commercially. For some bacterial species there are strains lacking some of the restriction systems. Treatment of the host with UV light or high temperatures has also been used to alleviate the effects of restriction systems (although the details of why this should be effective are not always clear). In addition, artificial methylation or demethylation may be used to protect against enzymes that are active on non-methylated or methylated DNA respectively. We will now consider the different methods for the introduction of DNA into recipient cells.

1. **Introduction of naked DNA.** Many bacterial species can be induced to take up naked DNA by suitable treatments, such as the ice-cold calcium solutions or electroporation used with *E. coli*. For a species that has not previously been studied, trial and error will usually be needed to optimize the methods.

   A number of genera, including the important Gram-positive *Bacillus*, *Streptococcus* and *Streptomyces*, include species that exhibit natural competence. That is, they can take up exogenous DNA without the need for non-physiological treatment, although the ocurrence and efficiency of natural competence can vary even within species. Competence is usually regulated, and it is often determined by the excretion into the growth medium of extracellular, low-molecular-weight proteins called competence proteins. Competence then develops as cell density (and the concentration of competence proteins) in the medium increases. The requirements for transforming DNA in many of these species are complex. For example, with *Bacillus*, plasmids that are capable of direct transformation into a host generally require sequences that are also present on the host's chromosome, or sequences that are also present on another plasmid already within the cell, or at least partial internal duplication of the plasmid sequence. It is probable that entry of the molecule into the cell is associated with cutting and degradation of one strand. Regeneration of a double-stranded molecule that can be replicated may then depend on an inter- or intra-molecular recombination event. If recombination occurs between the incoming DNA and a sequence in the host chromosome or an endogenous plasmid, then the result is integration of the DNA into that chromosome or plasmid. If the incoming DNA recombines with itself in an intramolecular

reaction, then the result is recircularization. In addition, for some species the presence of a specific short nucleotide sequence on the incoming plasmid assists (and in some cases is essential for) its uptake.

2. **Transformation of protoplasts.** For several groups of bacteria, especially the Gram-positive *Bacillus* and *Streptomyces*, the constraints on the types of plasmid that can be taken up by natural transformation can be relieved by use of protoplasts (cells from which the walls have been removed by treatment with lysozyme) in the presence of a suitable osmotic buffer, to stop cell lysis, and polyethylene glycol. (Some protocols use other reagents in addition to polyethylene glycol.) Normal cells are then regenerated from protoplasts by transfer to a suitable medium. The use of protoplasts somehow bypasses the stages that cause the linearization and partial degradation of the incoming DNA. In some cases it is possible to use liposomes or, indeed, protoplasts of related species to transfer material. Disadvantages of protoplast transformation include the difficulty of regenerating normal cells from protoplasts for some species and the fact that the rich media used for this may preclude the selection of markers (especially nutritional ones).
3. **Conjugation**. In conjugation, two cells become physically linked by a proteinaceous tube (or pilus) and DNA is passed from one cell to the other. This is how the F (for 'fertility') plasmid, which forms the basis of a lot of *E. coli* genetics, is transferred from one cell to another. The F factor encodes all the functions necessary for its own transfer. However, other, smaller, plasmids exist that are more suitable for artificial transfer between species. These are the broad host-range plasmids referred to above. The plasmid to be transferred (sometimes called the cargo plasmid) is often unable to direct its own transfer. Instead, it relies on a conjugal plasmid that is able to do so, and sometimes one or more helper plasmids that encode functions such as methylases that will protect the cargo from degradation in its future host. The helper plasmid may also carry functions needed for the transfer of the cargo. Typically, the transfer is achieved in a triparental mating (Figure 9.3), of the following three conjugants: *E. coli* carrying the conjugal plasmid; *E. coli* carrying the cargo and any helper plasmids; and the recipient bacterial species to be manipulated. Conjugation between the two strains of *E. coli* places all the plasmids necessary in one *E. coli* cell, and conjugation with the recipient bacterium transfers the cargo into the recipient species. There are variations on this approach. If all the plasmids can be assembled in one *E. coli* strain first, then a biparental mating is sufficient to introduce the plasmid of interest into the recipient strain. Although *E. coli* can conjugate with a wide range of bacteria, including some Gram-positive bacteria, other bacterial species can be used as a donor. If the same species can be used as donor and recipient, then

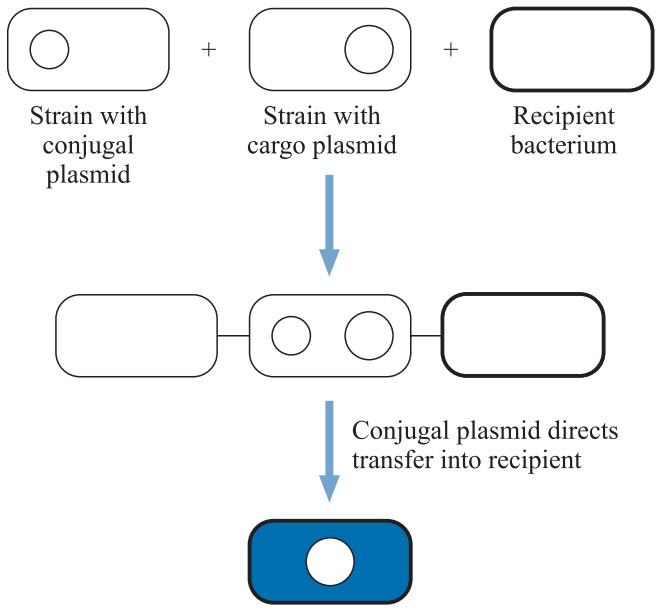


Fig 9.3 Triparental mating to mobilize a cargo plasmid into a recipient.

the plasmid transfer may be more efficient. Conjugation is often useful where direct transformation and protoplast transformation have failed.

9.2.3 Hosts

The same considerations for host design in *E. coli* apply to other bacteria. *B. subtilis* probably offers the best range of host strains. In this species, in addition to host nutritional markers, such as *trpC2*, *leuB6* and *metB10* (blocking tryptophan, leucine and methionine biosynthesis respectively), there are mutations available (such as *recE4)* that lead to reduction in recombination and plasmid instability. There are also mutations reducing endogenous restriction enzyme activity and protease activity (which may be useful when expressing cloned genes). Useful mutant strains are available for some other species, although genes such as *recA* that can be inactivated in some species may be essential (or very nearly essential, so that inactivation leads to very low viability) in others.

9.3 *Saccharomyces cerevisiae*

The fungi are an extremely diverse group of organisms. Perhaps the best studied is the ascomycete fungus *Saccharomyces cerevisiae* (also known as bakers' yeast, or simply yeast), so we will concentrate on that. However, vectors and transformation systems have been

Table 9.4. Selectable markers for use in *Saccharomyces cerevisiae*

Marker	Function
URA3	Orotidine-5′-phosphate decarboxylase (in *de novo* synthesis of pyrimidines)
LEU2	β-Isopropylmalate dehydrogenase (in leucine biosynthesis)
HIS3	Imidazole glycerol phosphate dehydratase (in histidine biosynthesis)
ARG4	Argininosuccinate lyase (in arginine biosynthesis)
TRP1	*N*-(5′-Phosphoribosyl) anthranilate isomerase (in tryptophan biosynthesis)
TUN^R	UDP-*N*-acetylglucosamine-1-P transferase (in glycosylation; confers tunicamycin resistance)
TCM1	Ribosomal protein (confers resistance to trichodermin)
CYH2	Ribosomal protein (confers resistance to cycloheximide)
ADE2	Phosphoribosyl amino-imidazole carboxylase (in *de novo* biosynthesis of purines)
CAN1	Permease (confers sensitivity to canavanine)
SUP4	Tyrosine-tRNA (ochre chain termination suppressor)

developed for a number of other fungi, including species that are of particular interest as pathogens for plants and animals, so we will look briefly at some of those as well. We will look at the markers available, the vectors containing them and the ways of getting DNA into host cells.

9.3.1 Markers

One of the most commonly used selectable markers is the *URA3* gene. This encodes orotidine-5′-phosphate decarboxylase, an enzyme of uracil biosynthesis. Acquisition of a functional *URA3* gene by cells that were previously mutant at this locus allows them to grow in the absence of exogenous uracil. Examples of other frequently used selectable markers, including nutritional markers that can be selected in a similar way, are given in Table 9.4. Some of the markers (such as *CAN1*, *SUP4* and *URA3*) can be counterselected, i.e. we can select for their absence. The *CAN1* gene encodes a permease that causes uptake of the toxic arginine analogue canavanine and consequent cell death. So, growth of cells in the presence of canavanine imposes selection for those lacking a wild-type *CAN1* gene. This also forms the basis of selection for *lack* of *SUP4*, using a host with an ochre chain termination mutation in the *CAN1* gene. The *can1* mutant host cells are resistant to canavanine, as they are unable to import it. However, the chain termination suppressor *SUP4* causes the production of functional permease in the *can1* background (because it suppresses the *can1* mutation) and, therefore, causes the death of *can1* cells in the presence of canavanine. Loss of *SUP4* abolishes canavanine uptake and causes canavanine resistance, which can readily be selected. Cells containing *SUP4* can also be identified (without killing them) by the use of hosts with an ochre chain termination mutation in the *ADE2* gene for phosphoribosyl amino-imidazole carboxylase, a component of the

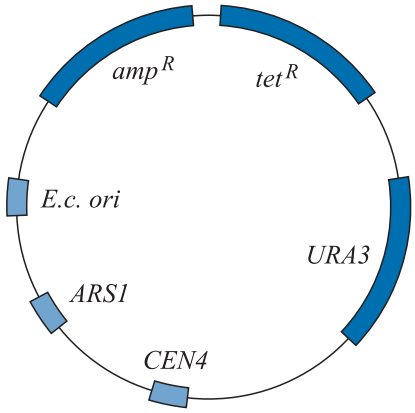


Fig 9.4 Yeast centromeric plasid YCp50 (7.9 kb). The presence of a suitable origin (*E. c. ori*) and selectable markers (amp^R and tet^R) allows it to be used as a shuttle vector in *E. coli*. The vector contains a yeast selectable marker (*URA3*), a centromeric sequence (*CEN4*) and an autonomously replicating sequence (*ARS1*).

purine biosynthesis pathway. The mutation causes the cells to accumulate a red pigment, whereas colonies are the usual colour if the ochre mutation is suppressed by *SUP4*. Cells that have lost *SUP4*, therefore, acquire the red pigment and are visibly distinguishable from the others. Cells mutant in *URA3* can be selected using 5-fluoro orotic acid (5-FOA), which is turned into toxic products by the wild-type protein.

9.3.2 Plasmid vectors

Integration of non-replicating incoming circular or linear DNA into the yeast genome by homologous recombination can be used to bring about stable genetic modification. The need for integration will clearly make recovery of the incoming DNA much more difficult than if a replicating vector is used. Integration is the basis of gene disruption, widely used for inactivating genes in yeast (see Chapter 7).

The replicating plasmid vectors used commonly in yeast can be divided into two main categories: yeast centromeric plasmid (YCp) and yeast episomal plasmid (YEp) vectors. An example of a YCp plasmid is given in Figure 9.4. It contains prokaryotic ampicillin and tetracycline resistance genes and an origin of replication for *E. coli*, all in a region of the plasmid derived from pBR322. These will allow it to function as a shuttle vector, and sequences can be cloned into the ampicillin or tetracycline resistance genes. Many of the yeast vectors are based on pBR322 in this way, but other plasmids have also been used including the pUC and Bluescript species. There are also a selectable marker and an origin of replication for yeast, *ARS1*. (An ARS, or autonomously replicating sequence, confers on a plasmid the ability to replicate without integration into another replicon. There are several available from yeast.) The distinguishing feature of YCp plasmids is the presence of a chromosomal centromeric sequence, *CEN4* in this example. The presence of a centromeric sequence allows the plasmid to be partitioned by the spindle apparatus of the cell, in the same way that the endogenous chromosomes are partitioned. Although this also results in a large reduction of the copy number per cell, it confers a much greater stability during cell division, allowing

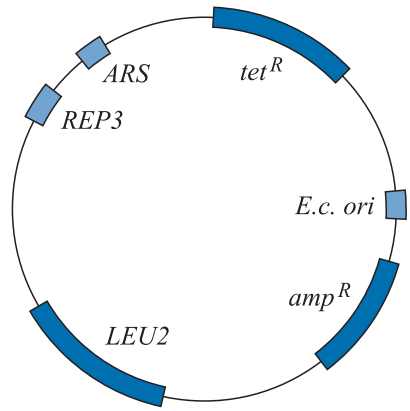


Fig 9.5 Yeast episomal plasmid YEpl3 (10.8 kb). The presence of a suitable origin (*E. c. ori*) and selectable markers (amp^R and tet^R) allows it to be used as a shuttle vector in *E. coli*. The vector contains a yeast selectable marker (*LEU2*), a partitioning sequence (*REP3*), and an autonomously replicating sequence (*ARS*) derived from the 2 μ plasmid.

the plasmid to be maintained in the absence of selection. Plasmids carrying an *ARS* but no *CEN* sequence are sometimes called YRp (yeast replicating plasmid) vectors. Although they have higher copy numbers than YCp vectors, their instability is a major drawback.

An example of a YEp vector is shown in Figure 9.5. Most of the features are similar to those of YCp plasmids, except for the absence of the centromeric sequence and its replacement with a different origin of replication. This comes from a naturally occurring episome (plasmid) from yeast, called the 2 μ (micron) circle. As well as the origin of replication, there is a naturally occurring *cis*-acting sequence, *REP3*, which is the site of action for proteins that help in the partitioning of the plasmid at cell division. This gives the YEp vectors a mitotic stability similar to that of the YCp vectors, but with a somewhat higher copy number. It may be possible to increase copy number further by use of a selectable marker (e.g. some alleles of *LEU2*) that complements only poorly. This gives a selective advantage to cells with a higher plasmid copy number.

There are more sophisticated versions of these classes of vector, with further modifications. These include the presence of filamentous phage origins for generation of single-stranded DNA and the *lacZ′* system for detecting the presence of inserts when the vectors are used in *E. coli*.

9.3.3 YACs

YACs are sophisticated cloning vectors that can be used for propagating large stretches (in the megabase size range) of DNA. This is very useful, because it reduces the number of recombinants needed to cover the entire genome of an organism. The aim is to construct an artificial chromosome that can be maintained in yeast, but which is mostly composed of foreign DNA. An example is shown in Figure 9.6 together with a representative cloning strategy. The vector contains *HIS3*, *URA3*, *TRP1* and *SUP4*, described above. *CEN4* is a centromeric sequence and *TEL* is a telomeric sequence derived from the ends of ribosomal RNA-encoding molecules from the macronucleus of the protozoan *Tetrahymena*. In addition, there

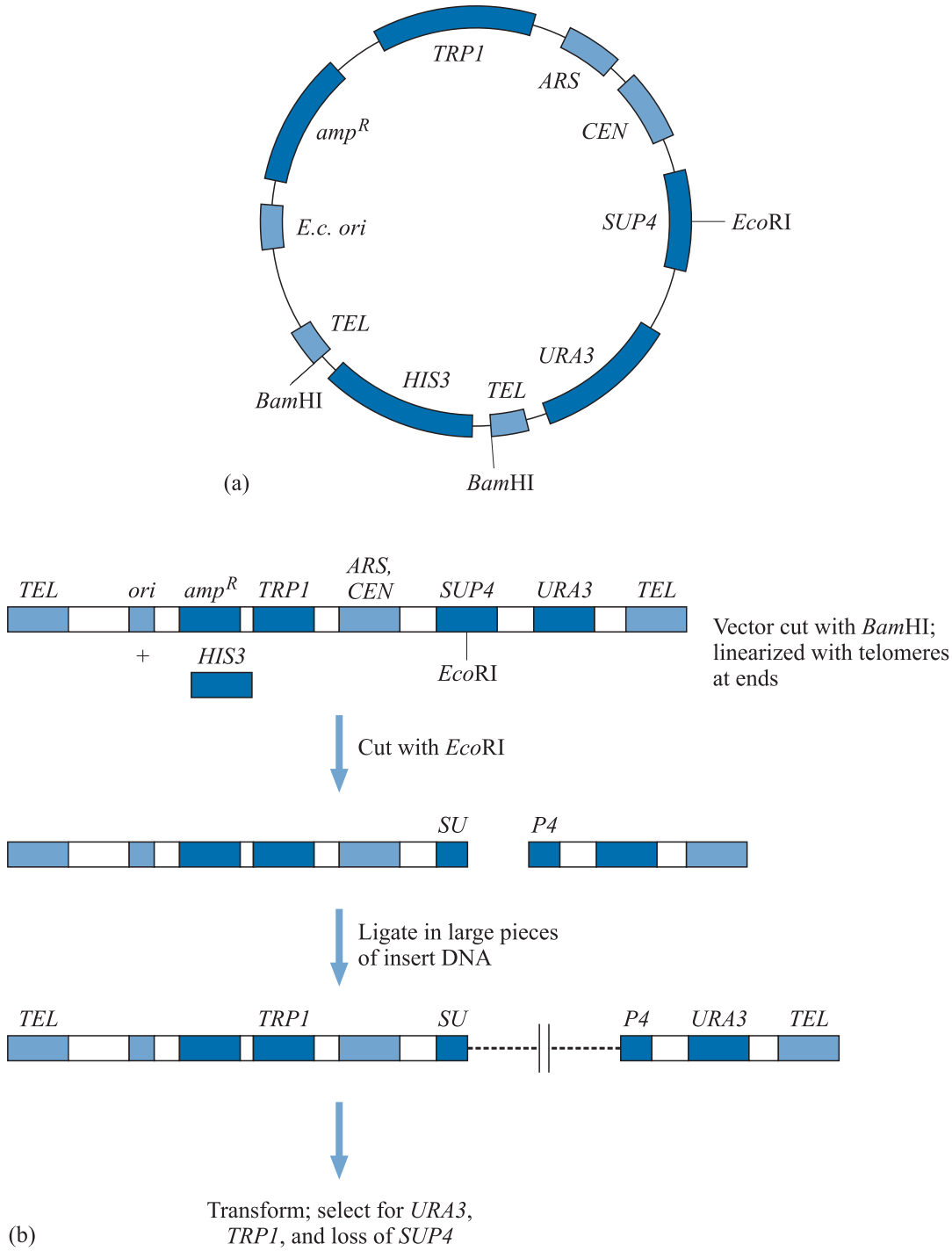


Fig 9.6 YAC cloning: **(a)** shows a representative YAC, pYAC4; **(b)** a typical cloning strategy. Note the telomeres (*TEL*) that constitute the ends of the linear recombinant molecule, as shown in panel **(b)**. Other sequences are as described in the text. The inserted DNA is indicated by the broken dashed line.

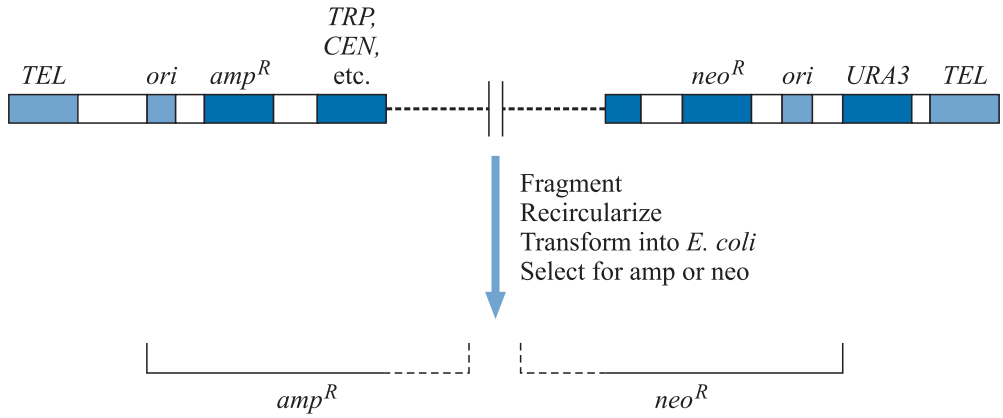


Fig 9.7 YAC cloning and recovery in *E. coli*. Note the presence of a prokaryotic origin of replication (*ori*) and drug resistance gene (amp^R or neo^R, for resistance to ampicillin and neomycin) in each arm, allowing recovery of each arm and adjacent inserted sequences.

are a yeast origin of replication, a prokaryotic origin of replication, and an ampicillin resistance gene. The latter two features allow YAC DNA to be prepared from *E. coli* prior to the cloning. The whole molecule is just over 10 kb.

The YAC DNA is digested, in this case with *Bam*HI and *Eco*RI; this generates three fragments, two of which contain *TEL* sequences and which are the ones we will need. (The third has the *HIS3* gene.) Subsequent religation of the fragments can be prevented by treatment with phosphatase. The DNA is then mixed with the insert material (which has been prepared in a way that leaves *Eco*RI ends) and ligated and the ligation products are introduced into yeast. There, they can function as *bona fide* chromosomes, possessing a centromere, an origin of replication and telomeres. Transformed cells can be selected by the presence of the *URA3* and *TRP1* genes. The presence of both of these markers indicates that both arms of the YAC have been acquired. Screening for the inactivation of *SUP4* allows putative recombinants to be identified.

There are several modifications to this basic YAC design. One is the incorporation of an extra prokaryotic origin and prokaryotic selectable marker, so that there is one of each on each arm. This assists in the recovery of at least parts of individual recombinant YACs using *E. coli*. The YAC DNA is isolated from yeast, digested, self-ligated and used to transform *E. coli*. Cells containing either the left arm or the right arm can then be selected (Figure 9.7). Bacteriophage promoters (e.g. from T7) can also been included to allow the synthesis of RNA transcripts in vitro.

9.3.4 Expression systems

A number of constitutive promoters are available, e.g. from the *GPD* gene for glyceraldehyde-3-phosphate dehydrogenase, *TEF* for translation elongation factor 1α, and *CYC1* for a component of cytochrome *c*

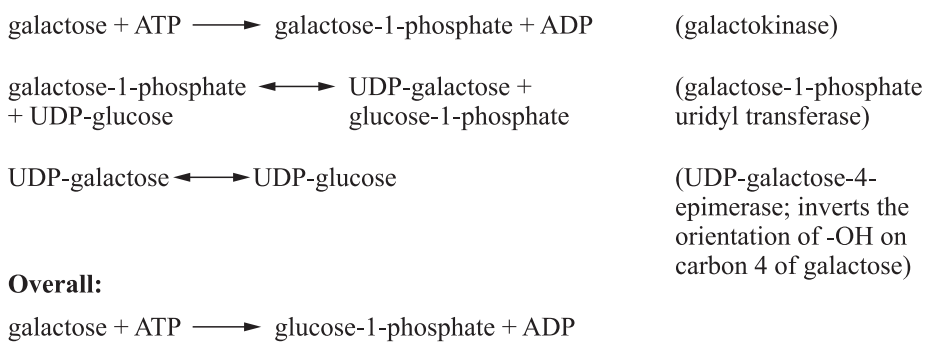


Fig 9.8 Metabolism of galactose. The glucose-1-phosphate can be converted to glucose-6-phosphate through glucose-1,6-bis-phosphate and then metabolized by glycolysis or other standard routes.

Table 9.5. Regulatable promoters for use in *Saccharomyces cerevisiae*

Promoter	Function and control
GAL1	Galoctokinase; upregulated in presence of galactose, repressed in presence of glucose
GAL10	UDP-galactose-4-epimerase; upregulated in presence of galactose, repressed in presence of glucose
PGK	3-Phosphoglycerate kinase; upregulated in presence of glucose, repressed on non-fermentable carbon sources, such as acetate
ADH1	Alcohol dehydrogenase; upregulated in presence of glucose, repressed on non-fermentable carbon sources, such as acetate
PHO5	Secreted acid phosphatase; upregulated in medium deficient in inorganic phosphate (to assist scavenging), repressed in presence of phosphate
MET25	O-Acetyl homoserine sulphydrylase (enzyme of methionine biosynthesis); repressed in presence of methionine
CUP1	Copper detoxification; upregulated when copper present in growth medium
CTR1, CTR3	Copper transport; upregulated when copper absent from medium

oxidase. Some of the most widely used controllable promoters come from genes for galactose metabolism (Figure 9.8), *GAL1* and *GAL10*, encoding galactokinase and UDP-galactose-4-epimerase respectively. When wild-type control proteins are present, transcription of *GAL1* and *GAL10* is increased greatly in the presence of galactose. In the presence of glucose, transcription is tightly repressed. There are two deletion variants of the *GAL1* promoter, designated *GALL* and *GALS*, that are less strong and, therefore, less likely to lead to problems caused by toxicity of foreign proteins. All these promoters, together with their upstream controlling sequences, can be used either in expression vectors or as cassettes for insertion into other constructs. These and other promoters that can be tightly controlled and are used for expression purposes are summarized in Table 9.5. All these promoters can be used to direct the synthesis of either unmodified proteins or fusion proteins; the latter can incorporate

either the corresponding products of *GAL1*, *GAL10* and so forth, or other sequences, such as LacZ or TrpE from *E. coli*. An advantage with using yeast rather than prokaryotic cells for expression of eukaryotic proteins is that post-translational modifications of the proteins (such as glycosylation, phosphorylation, and acylation) are unlikely to take place in prokaryotic systems but they may do so in yeast. Even in yeast, these modifications may not take place reliably, and it may be necessary to use expression in insect or mammalian cells (see below).

9.3.5 Secretion systems

Secretion of overexpressed proteins by fusion to a secretion signal may be useful because it assists purification and allows the proteins produced to be only minimally exposed to most of the yeast intracellular proteases. In addition, it may help to bring about correct glycosylation and may also enhance protein folding. A number of secretion signals, from both yeast and other organisms, can be used. Those for the mating pheromone alpha-peptide and for invertase (product of the *SUC2* gene) are often used. The mating pheromone is a 13-residue peptide, secreted from cells of alpha mating type, that acts on a receptor on cells of mating type a. Invertase is secreted in order to bring about the breakdown of sucrose to glucose and fructose.

9.3.6 Transformation systems

Three ways are commonly used for introducing exogenous DNA into *S. cerevisiae*. One of the simplest and most widely used is the treatment of cells with lithium acetate and polyethylene glycol and a heat shock, together with the transforming DNA and additional non-specific denatured 'carrier' DNA (e.g. from calf thymus). This method is relatively fast and straightforward and can give an efficiency of transfer of up to 10^5 to 10^6 colonies per microgram of plasmid with suitable strains.

A second approach to introducing exogenous DNA into *S. cerevisiae* is the transformation of protoplasts (also called spheroplasts), which are obtained by digestion of the cell wall with appropriate enzymes, in the presence of a suitable osmotic buffer, such as sorbitol solution, to prevent lysis. The protoplasts are exposed to the transforming DNA, in the presence of polyethylene glycol and calcium chloride. The next step is to regenerate intact cells from protoplasts. This is achieved by immobilizing the protoplasts in a suitable agar that is osmotically buffered. Surrounding the protoplasts with a solid medium allows the retention of cell wall material around the cell during the earliest and most sensitive stages of cell wall regeneration. Protoplast transformation is more time consuming than transformation with lithium acetate, but with some strains it can give a higher transformation efficiency. Therefore, it is useful

with strains that otherwise have a low transformation efficiency, or when the amount of DNA available is low.

A third method for introduction of DNA into *S. cerevisiae* is electroporation. It is reported that this method can give even higher efficiency than protoplast transformation, although the actual figure seems to vary quite widely and probably depends on the strain used. A disadvantage is that the number of transformants obtained reaches a plateau with relatively low amounts of DNA. As with other electroporation systems, the cells are simply subjected to an electric field pulse. There is no need to make protoplasts for this treatment, although an osmotic buffer such as sorbitol is often included. Treatment with lithium acetate and dithiothreitol is also sometimes included to increase transformation efficiency.

Other less widely used methods for transformation of *S. cerevisiae* include particle bombardment and cell damage by agitation with glass beads. Transmission of DNA from *E. coli* by conjugation has also been reported. The *E. coli* cells must contain a plasmid that can replicate in *E. coli* and in yeast (i.e. a shuttle vector) and another plasmid to provide mobilization functions.

9.4 Other fungi

9.4.1 *Schizosaccharomyces pombe*

Another unicellular fungus that is frequently studied is the fission yeast *Schizosaccharomyces pombe*. It is particularly widely used in work on the cell cycle and eukaryotic chromosome biology. As its name implies, *Schizosaccharomyces pombe* reproduces by fission, in contrast to *Saccharomyces cerevisiae*, which reproduces by budding. Many of the early vectors for use with *Schizosaccharomyces pombe* contained the origin of replication from the 2μ plasmid of *Saccharomyces cerevisiae*. However, this does not always allow stable maintenance in *Schizosaccharomyces pombe*, and a more common strategy now is to use plasmids with a *Schizosaccharomyces pombe* origin of replication, *ars1*. However, because the *Schizosaccharomyces pombe* centromere is very large, it is not possible to include it in such plasmids. Consequently, the copy number of these plasmids can vary widely and the plasmids may be lost altogether. Therefore, it may be better to direct the integration of incoming DNA, which can be achieved either by homologous or by non-homologous recombination.

A number of selectable markers are available. One of the most widely used is *ura4*$^{+}$, which encodes orotidine-5′-phosphate decarboxylase (the enzyme encoded by *URA3* of *Saccharomyces cerevisiae*). This can be selected in a *ura4*$^{-}$ background in the usual way. (The *URA3* gene of *Saccharomyces cerevisiae* is sometimes used, but is less efficient.) Absence of *ura4*$^{+}$can be selected using 5-fluoro-orotic acid, which is converted to a toxic product by the

enzyme. The *leu1*$^+$ marker, encoding β-isopropylmalate dehydrogenase, is also popular. As with *ura4*$^+$, the *Saccharomyces cerevisiae* equivalent (*LEU2*) is also used, but is less efficient. The *sup3–5* gene of *Schizosaccharomyces pombe* is often used, as a suppressor of a mutation in the *ade6* locus (equivalent to *ADE2* of *Saccharomyces cerevisiae*). Markers based on genes for histidine biosynthesis (*his3*$^+$ and *his7*$^+$), arginine biosynthesis (*arg3*$^+$) and antibiotic (G418 or hygromycin) resistance are also available. There are counterselection systems for loss of the *Saccharomyces cerevisiae CAN1* gene and a thymidine kinase gene from herpes simplex virus, as well as the *ura4*$^+$gene. The replacement of one plasmid by another by introduction of the second plasmid into a strain containing the first, followed by removal of the first by counterselection is sometimes called a plasmid shuffle.

Expression vectors are available with a range of constitutive or inducible promoters. Constitutive promoters include that from the *adh1*$^+$ alcohol dehydrogenase gene, as well as those for the small-T antigen transcript of SV40 and the 35S transcript of the plant virus CaMV. (The last of these has been modified to contain a tetracycline control system derived from a bacterial transposon, so that expression is induced in response to added tetracycline.) One of the most widely used inducible promoters is from the *nmt1*$^+$ gene. This gene presumably encodes a polypeptide involved in thiamine bioysnthesis, as there is *n*o *m*essage in *t*hiamine. There are various forms of this promoter, which have been artificially modified to provide a range of promoter strengths. The promoter from the *fbp1*$^+$ gene for the gluconeogenic enzyme fructose 1,6-bis-phosphatase is also widely used. The promoter is tightly repressed in the presence of glucose, but activated when glucose is absent and glycerol is present as a carbon source instead (conditions when the gluconeogenesis pathway is active). Other regulated promoters include those for invertase (induced by growth in the presence of sucrose and absence of glucose) and a copper transporter (induced by growth in the absence of copper).

Early work on *Schizosaccharomyces pombe* used protoplast transformation for the introduction of DNA. However, this has now largely been replaced by electroporation and transformation using lithium acetate. This is probably because protoplast transformation is more time consuming, and it is also reported to favour non-homologous integration when using non-replicating sequences.

9.4.2 Other yeasts

Other yeasts may be useful in having more efficient secretion systems than *Saccharomyces* and *Schizosaccharomyces*. These include *Kluyveromyces lactis*, *Hansenula polymorpha* and *Pichia pastoris*. There are integration vectors available for *Kluyveromyces lactis*, and also vectors for extrachromosomal replication using chromosomal ARSs or origins from naturally occurring *Kluyveromyces* plasmids.

Secretion can be directed in *Kluyveromyces* by fusion to suitable sequences from a secreted toxin that kills other cells. The yeasts *Hansenula polymorpha* and *Pichia pastoris* are methylotrophic. That is, they can grow on methanol as a carbon source, and this induces the expression of genes for enzymes such as methanol oxidase (also known as alcohol oxidase) and formate dehydrogenase. This induction has been used as the basis for expression vectors. Proteins can be secreted at high levels using a range of signal sequences, including the signal sequence from the alpha mating factor of *Saccharomyces cerevisiae*, and there is generally little proteolysis of secreted material. The methanol oxidation pathway depends on the peroxisome, and these yeasts have also been exploited for the study of targeting to the peroxisome.

9.4.3 Filamentous fungi

There is a lot of interest in the transformation of other fungi (and other organisms loosely classified as fungi, such as oomycetes). Filamentous fungi are of particular importance, as they include a wide range of plant and animal pathogens and biological control agents. Techniques for the genetic manipulation of the model filamentous fungi *Neurospora crassa* and *Aspergillus nidulans* were developed not long after those for *Saccharomyces cerevisiae*. Methods are now available for a wide range of such organisms, including *Acremonium chrysogenum* (used for cephalosporin production), *Aspergillus oryzae* (used in a range of fermentations), *Fusarium oxysporum* (a pathogen with a broad host range), *Magnaporthe grisea* (causative agent of rice blast disease), *Penicillium chrysogenum* (used for penicillin production), *Trichoderma viride* (a wood-digesting fungus also used for biological control), *Ustilago maydis* (causative agent of smut disease on maize) and the oomycete genus *Phytophthora* (including the pathogen responsible for the Irish potato famine in the mid-19th century).

A wide range of selectable markers is available. They are summarized in Table 9.6. Some of these require isolation of an appropriate mutant strain in which to select the marker. Strains deficient in orotidine-5′-decarboxylase can be isolated by virtue of their resistance to 5-fluoro-orotic acid, and strains deficient in nitrate reductase can be isolated by their resistance to chlorate (which generates a toxic product under the action of nitrate reductase). Many markers, including their promoters, can be used across a wide range of species. So, for example, the $argB^+$ gene from *Aspergillus nidulans* was used in early work on *Magnaporthe grisea*, and the $niaD^+$ gene of *Aspergillus nidulans* was used in a $niaD^-$ strain of *Fusarium oxysporum*. Many promoters (such as heat shock promoters) and origins of replication function effectively in a wide range of fungi. The origin of replication of the *Saccharomyces cerevisiae* 2 μ plasmid can also be used in a number of filamentous species. In some cases, stable maintenance of extrachromosomal molecules is

Table 9.6. Selectable markers commonly used in filamentous fungi

Marker	Source	Product	Phenotype
argB	*Aspergillus nidulans*	Ornithine trancarbamoylase	Ability to synthesize arginine
hph	*Escherichia coli*	Hygromycin phophotransferase	Hygromycin resistance (phosphorylates it)
ble	*Streptoalloteichus hindustanus*	Bleomycin-binding protein	Bleomycin resistance (binds to it and inhibits its DNA-damaging activity)
pyrG	*Aspergillus nidulans*	Orotidine-5′-decarboxylase	Ability to synthesize uracil (= *URA3* of *Saccharomyces cerevisiae*, and *pyr-4* of *Neurospora crassa*)
niaD	*Aspergillus nidulans*	Nitrate reductase	Ability to grow with nitrate as sole nitrogen source
benA	*Aspergillus nidulans*	Beta-tubulin	Resistance to fungicide benomyl
amdS	*Aspergillus nidulans*	Acetamidase	Growth on acetamide as sole carbon or nitrogen source

difficult, because of the filamentous nature of the fungi (so that cells that have lost a selectable marker can be maintained by those retaining it), so integration is more convenient. DNA may be integrated into host chromosomes by homologous or non-homologous recombination.

There are a number of methods for the introduction of DNA. In early work, transformation was carried out with protoplasts made from spores or hyphae by digestion of the cell wall with suitable enzymes. DNA is added in the presence of calcium ions and polyethylene glycol. The latter apparently causes protoplast fusion, and DNA is taken up at the same time. The protoplasts are then regenerated. Another method is to expose germinating spores to lithium acetate and polyethylene glycol together with the transforming DNA. This circumvents the need for making and regenerating protoplasts. Electroporation is also commonly used, either with protoplasts or with other cell types. Where cells have thick walls and protoplast production is difficult, biolistic transformation is often useful. In many cases, if integration of a marker is selected, it is found that other, non-selected, DNA species that were applied at the same time are also integrated. This is called co-transformation. It seems likely that a subset of the cells in the experiment is able to integrate incoming DNA, but does so at high efficiency. In effect, the selected marker allows this subset to be identified, and is highly likely to have integrated the other DNA species at the same time.

9.5 Algae

The unicellular green alga *Chlamydomonas reinhardtii* has been studied for many years, both as an important organism in its own right and as a model system for plants. It is particularly attractive in allowing transformation of the chloroplast and mitochondrial genomes in addition to the nuclear genome reference to 9.7 (Organelle transformation). Methods are also being developed for the manipulation of other algal species.

9.5.1 *Chlamydomonas reinhardtii*

The yeast *ARG4* gene for argininosuccinate lyase (an enzyme of arginine biosynthesis, converting argininosuccinate to arginine and fumarate) was used as a selectable marker in early work with *Chlamydomonas*. However, the *Chlamydomonas* genome is very GC rich. This means that many heterologous genes are not well expressed, unless they come from genomes that are also GC rich. The use of the yeast *ARG4* gene, therefore, was superseded by its *Chlamydomonas* homologue, *ARG7*. The most frequently used markers are listed in Table 9.7. Several require a suitable mutant background for selection. However, a number confer resistance to antibiotics or other toxic compounds and can be selected in a wild-type background. Other markers that are less widely used include genes for components of the flagellum and the *OEE1* gene, which encodes a component of Photosystem II and allows mutant cells lacking that polypeptide to grow photoautotrophically.

Although *ARS* sequences have been obtained from *Chlamydomonas*, most experiments simply use integration of DNA into the chromosome, which gives efficient transformation. Integration is usually by non-homologous recombination. This means that gene disruption (see Chapter 7) is difficult to achieve. However, RNAi can be used to reduce the level of expression of a target gene.

Transformation of *Chlamydomonas* is most efficient if protoplasts are used. These can be generated by treatment of intact cells

Table 9.7 Selectable markers commonly used in *Chlamydomonas reinhardtii*

Marker	Product	Properties
ARG7	Argininosuccinate lyase	Arginine biosynthesis (replaced *Saccharomyces cerevisiae ARG4*)
NIA1	Nitrate reductase	Growth on nitrate as sole nitrogen source (sensitivity to chlorate)
ble	Bleomycin-binding protein	Bleomycin (phleomycin, zeomycin) resistance
ALS	Acetolactate synthase	Resistance to sulphonylurea
aadA	Aminoglycoside adenyltransferase	Resistance to spectinomycin
aph7	Aminoglycoside phosphotransferase	Resistance to paromomycin, hygromycin B (also *aphVIII*)
MAA7	Tryptophan synthase	Counterselectable with 5-fluoroindole

with autolysin, a cell-wall digesting activity that is produced by gamete cells immediately prior to mating. Pre-incubation in medium lacking nitrogen can also be used to induce autolysin production. There are also mutant strains, most notably the one designated *cw15*, that lack a cell wall and hence form a 'natural' protoplast. DNA can be introduced into protoplasts by simple treatment with chemicals such as poly-L-ornithine and polyethylene glycol or zinc sulphate, but this often gives low and variable transformation rates. More efficient methods include agitation of cells in the presence of DNA and small glass beads or silicon carbide whiskers, biolistic transformation and electroporation. Transformation with *Agrobacterium* (see below) has also been reported. In general, electroporation gives the highest transformation efficiency, but agitation with glass beads is often used in preference because of its simplicity. Co-transformation is often seen.

9.5.2 Other algae

There are reasonably well-developed systems for the genetic manipulation of the green algae *Volvox carteri* and *Chlorella* sp., often using nitrate reductase or antibiotic resistance as selectable markers. Biolistic transformation has been used for both genera, and chemically induced DNA uptake by protoplasts has been used with *Chlorella*. There is considerable interest in developing effective methods for green algae of the genera *Haematococcus* and *Dunaliella*, which are of economic importance for the production of carotenoids. There are also several reports of transformation of red algae, such as *Cyanidioschyzon merolae*, and brown algae, such as the diatom *Phaeodactylum tricornutum* and the seaweed *Laminaria japonica*.

9.6 Vascular plants

A wide range of methods has been used for the transformation of plant cells. The most widely used model plant is *Arabidopsis thaliana*, though many other plants, such as tobacco, are also important as model systems. Transformation of economically important crop species, especially cereals, is also of particular interest. We will look at the different methods available, the most widely used of which is *Agrobacterium*-mediated transfer.

9.6.1 *Agrobacterium*-mediated transfer

This exploits the natural DNA-transferring properties of a soil-inhabiting bacterium, *Agrobacterium tumefaciens*. A related species, *Agrobacterium rhizogenes*, can also be used. The two species cause the diseases crown gall and hairy root respectively on plants. The species belong to the Rhizobiaceae, which include other important bacterial genera, such as *Rhizobium*. Both *Agrobacterium* species lead to disturbance of normal plant growth, causing a gall in one case

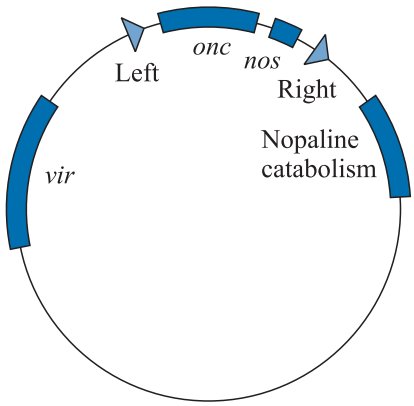


Fig 9.9 Simplified diagram of a Ti plasmid. Arrowheads indicate the left and right borders of the region transferred (the T-DNA). *Onc* genes are responsible for growth transformation, *nos* for nopaline synthase, and *vir* for transfer of the T-DNA into a target cell.

and abnormal roots in the other. Their ability to do so depends on the presence of large plasmids, in excess of 100 kbp, which are referred to as the 'tumour-inducing' (Ti) and 'root-inducing' (Ri) plasmids respectively. We will concentrate on the Ti plasmid; the Ri plasmid is quite similar.

A simplified diagram of a Ti plasmid is shown in Figure 9.9. Most important is a region referred to as the T-DNA ('transferred DNA'), which is transferred from an infecting *Agrobacterium* cell into the nucleus of the plant cell, where it is integrated into the plant genome. Transfer of the T-DNA depends on a set of genes called *vir* if they are on the Ti plasmid, or *chv* if they are on the chromosome. They are induced in response to various compounds in exudates from wounded plants; these compounds include acetosyringone, alpha-hydroxyacetosyringone and precursors of lignin, such as coniferyl alcohol. The T-DNA itself is flanked by repeated sequences of around 25 base pairs, called border repeats (or left and right borders). The T-DNA contains a group of genes referred to as the *onc* genes, as they are responsible for the T-DNA's oncogenicity (or tumour-inducing potential). They include genes for the synthesis of an auxin and a cytokinin, which are plant growth regulators that cause the disturbance of cell growth. In addition, the T-DNA contains genes for enzymes that produce unusual amino acid conjugates, such as nopaline and octopine (Figure 9.10), collectively called opines. Therefore, plant cells that have had T-DNA incorporated will grow in an apparently uncontrolled way and will synthesize opines. The opines are excreted by the plant cells (which are unable to use them themselves) and are utilized instead by *Agrobacterium* cells near the site of infection, whose Ti plasmids also contain the genes for opine breakdown. *Agrobacterium*, therefore, has a very sophisticated mechanism for subverting the normal cell function.

Vectors

The basic principle of using *Agrobacterium* as a tool for plant genetic manipulation is to insert foreign DNA into the T-DNA of a bacterial cell and rely on the bacterium to transfer the DNA into the plant.

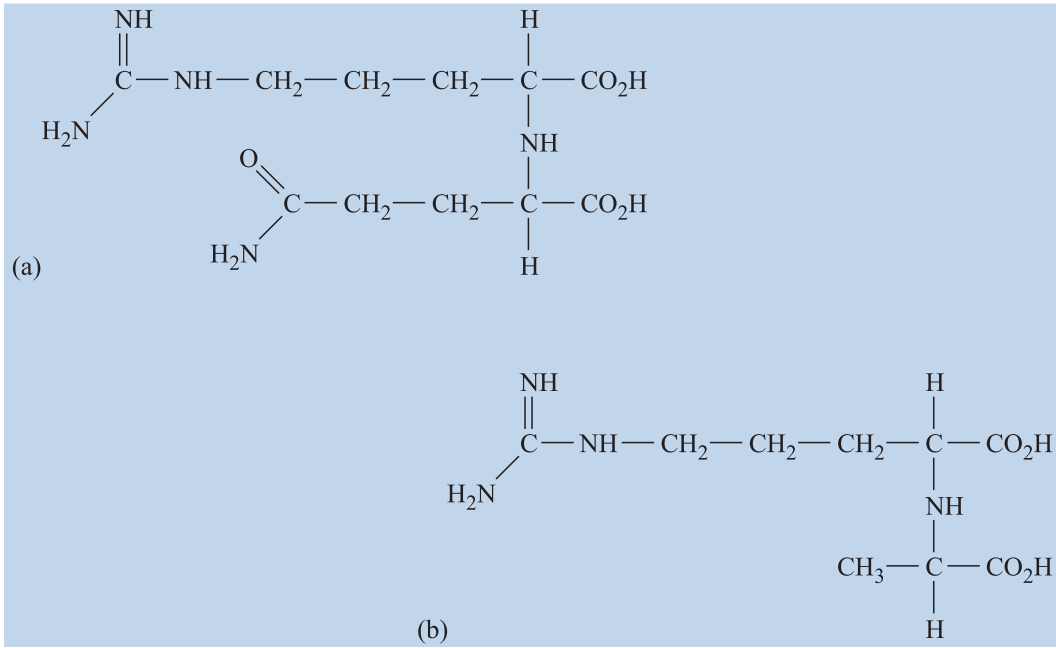


Fig 9.10 Examples of opines: (a) nopaline and (b) octopine.

As long as the necessary proteins are provided by the bacterium, any sequences flanked by the T-DNA border repeats can be transferred into the recipient plant cell genome.

There are two difficulties with this. The first is that the Ti plasmid is too large to manipulate in vitro. The second is that T-DNA contains genes that will disturb the growth of plants into which it has been transferred. These problems were initially overcome by the use of cointegrative vector systems. These used a small intermediate vector that could be manipulated conveniently and carried a selectable marker. The gene of interest was inserted into this and the recombinant intermediate vector introduced into *Agrobacterium* cells containing a Ti plasmid. The intermediate vector was not able to replicate inside *Agrobacterium*, but could be propagated by recombinational insertion into the T-DNA region of the resident Ti plasmid. (This event could be selected for using the marker on the intermediate vector.) *Agrobacterium* could then transfer the T-DNA into the recipient plant. For the recipient plants not to have abnormal growth, it was also necessary for the resident Ti plasmid to have been disabled, by inactivation of one or more of the genes for biosynthesis of growth regulators.

This approach was rather complex, because it required recombination between the integrative vector and the resident Ti plasmid in vivo in the *Agrobacterium* cell. It was superseded by the use of binary vector systems (Figure 9.11). These also use a small plasmid, the binary vector, that can be readily manipulated in vitro; but, unlike the cointegrative vector systems, this plasmid contains a pair of border repeats. The DNA of interest is inserted into the region of this plasmid between the border repeats. The recombinant binary

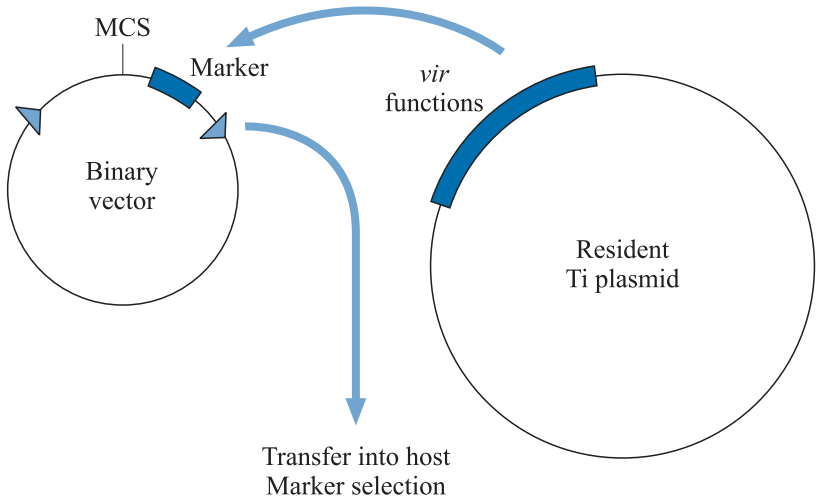


Fig 9.11 Binary vector systems. A disabled resident Ti plasmid directs transfer of T-DNA from a smaller vector (bounded by arrowheads in the figure) into the host genome. Selection is then imposed for the marker gene carried in the T-DNA.

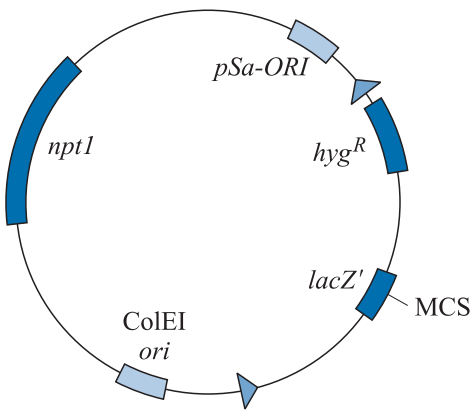


Fig 9.12 The binary vector pGreenII.000 (3.3 kb). The border repeats (arrowheads) flank T-DNA containing a *lacZ'* sequence with a multiple cloning site (MCS), and a selectable marker (hyg^R; hygromycin resistance). The pSa-ORI origin of replication and *nptI* marker allow propagation in *Agrobacterium*.

vector is then introduced into *Agrobacterium* containing a resident Ti plasmid (which has been disabled so that it cannot itself be transferred into a plant host). The *Agrobacterium* cells are then used to transfer DNA into the plant host. The resident Ti plasmid provides the functions needed (from its *vir* genes) to direct transfer of the region within the binary vector that is flanked by border repeats into the recipient plant genome.

The features of binary vectors are illustrated by the example shown in Figure 9.12. This particular one belongs to the pGreen series. It is a shuttle vector, so it can be propagated in *E. coli* or in *Agrobacterium*. In this case, there are two origins of replication. The ColEI origin allows replication in *E. coli* and the pSa-ORI allows replication in a suitable *Agrobacterium* strain. Some binary vectors use a single origin that can function in both hosts. The *nptI*

gene encodes a neomycin phosphotransferase that confers kanamycin resistance and allows the presence of the plasmid to be selected in a bacterial host. The left and right borders mark the region that will eventually be transferred to the plant host. Within this region are a multiple cloning site within a *lacZ′* minigene and a selectable marker (in this case for hygromycin resistance) allowing selection in the plant hosts. Selectable markers are discussed in more detail below. Other binary vectors include features such as additional plant promoters to drive expression and *cos* sites to allow packaging into lambda phage heads.

Once the correct sequences have been incorporated into the multiple cloning site of the binary vector, they must be transferred into a recipient *Agrobacterium* strain carrying an appropriate Ti plasmid. This is usually done by electroporation with naked DNA. However, transfer by conjugation can also be used. Conjugation often requires a triparental mating involving the *Agrobacterium* strain, an *E. coli* strain containing the vector to be transferred and an *E. coli* strain with a plasmid capable of mobilizing the binary vector into *Agrobacterium*.

Transfer into plant cells

The next stage is to arrange for the modified *Agrobacterium* to infect plant cells and transfer its T-DNA into them. Various methods exist, and the optimal ones vary from species to species. One approach (co-cultivation) is to incubate explanted material, such as leaf discs, shoots or roots, with *Agrobacterium* cultures containing the appropriate vector for a while and then transfer them to a suitable medium to select for transformation. This medium contains a selection system (according to which selectable marker was present in the vector), an additional antibiotic to kill residual *Agrobacterium* cells and plant growth regulators to assist in regeneration. Intact plants are then regenerated by tissue culture techniques, and these should contain the DNA transferred from the vector. An approach widely used with the model plant Arabidopsis is floral dipping. In this, plants are dipped briefly into a culture of *Agrobacterium* and then grown until seeds are produced. The seeds are subsequently germinated on an appropriate selective medium, and the seedlings that are able to grow should contain the DNA transferred from the binary vector. Spraying cultures onto floral tissues may also be adequate. In some protocols, plants are subjected to vacuum treatment while immersed in the bacterial culture. This is called vacuum infiltration. Other techniques for transfer of DNA from *Agrobacterium* include injection of bacteria into plant tissue and incubation of protoplasts with bacterial cultures followed by regeneration of plants from the protoplasts. In many cases the resulting plants can have several copies of the incoming DNA integrated at more or less random sites, sometimes with rearrangements, so detailed analysis of the transformed plants will be necessary.

9.6.2 Other methods of DNA transfer

Not all plant species can be successfully transformed using *Agrobacterium*. Another approach is to use biolistic transformation. This was first achieved using intact epidermal cells from onion as the target and microprojectiles carrying RNA from tobacco mosaic virus or a DNA construct encoding chloramphenicol acetyltransferase. Expression of both these nucleic acids was detected. In most applications, the gun is used to fire the microprojectiles into suitable target tissue such as leaf, after which transformed plants are regenerated on an appropriate selective growth medium.

Direct transformation of protoplasts without using *Agrobacterium* is possible. Uptake of naked DNA can be brought about by treatment with polyethylene glycol or by electroporation. DNA is integrated into the recipient genome, and in some cases at least this appears to be brought about by homologous recombination. Selection for transformants can be done using the same selectable markers as are used in *Agrobacterium* transformation. However, the most difficult aspect of this approach is the regeneration of intact plants from protoplasts. Although regeneration is possible for many plants (unlike most other eukaryotes), for many agriculturally important plants it is very difficult.

Other techniques are available, although they are less widely used than biolistic transformation or protoplast transformation. One technique uses viruses such as cauliflower mosaic virus (CaMV). Cloned DNA from CaMV is infectious when rubbed onto leaves of susceptible plants. Therefore, DNA can be inserted into the CaMV genome and then transferred into plants by rubbing onto the leaves. Although there are significant restrictions as to where DNA can be inserted into the CaMV genome to retain viability, modifications to at least one of the genes, ORF II, are possible. In one example, a chimeric molecule containing a sequence encoding a methotrexate-resistant version of the enzyme dihydrofolate reductase was introduced into turnip plants. The resulting plants produced the methotrexate-resistant enzyme and showed increased tolerance of the inhibitor when it was sprayed onto their leaves. Although this demonstrated the feasibility of the approach under certain circumstances, it has not found general application; this is partly because of the narrow constraints on the sequences that the virus can accept without loss of viability and partly because the viral DNA is presumably not integrated into the recipient, and transformation is not stable. Other methods that have been tried include direct injection of DNA, fusion of protoplasts with DNA-containing liposomes, biolistic transformation of pollen (followed by pollination with the transformed pollen) and injection of DNA into pollen tubes. These approaches are much less widely used, and in some cases there may be doubt as to whether transformation has actually taken place.

9.6.3 Transient expression

For some experiments it may not be necessary to regenerate intact plants. For example, many studies on the regulatory elements necessary for the expression of particular genes have been carried out by introducing the elements (or modifications of them) into protoplasts without stable transformation and then monitoring expression. Because the transformation is not stable, these are called transient expression systems (although, if things are working well, expression can be followed for days). When interpreting the results of these experiments, it should always be borne in mind that expression in protoplasts can be rather different from expression in the intact tissue, but for many situations the results appear quite reliable. The particle gun can be used for transient expression in target tissue that is more intact than protoplasts.

9.6.4 Promoters and selectable markers

One of the most widely used promoters is from the CaMV. It is called the 35S promoter, as it is responsible for the generation of a 35S transcript in infected cells. It is constitutive and very active, as is the *nos* promoter, which is also widely used. This comes from the nopaline synthase gene from the Ti plasmid. Although the *nos* gene is prokaryotic, it is functional in plants, where it directs the synthesis of the opine, which is exploited by other *Agrobacterium* cells. Slight adjustments have been made to the sequence to improve expression. Both the 35S promoter and the *nos* promoter have relatively little tissue specificity in their expression. Tissue specificity can be achieved by incorporation of control sequences from other promoters, such as the *rbcS* promoter (for the gene encoding the small subunit of ribulose bis-phosphate carboxylase), which is particularly active in illuminated leaves and stems. The *rbcS* promoter is also often used without association with the 35S or *nos* promoters. A very wide range of promoters is now available, to direct expression in many different tissue types. A tetracycline-regulated system (see Section 9.10.1) can also be used, directing expression in response to added tetracycline analogues.

The majority of selectable markers used confer resistance to antibiotics or herbicides. The most widely used selectable marker is resistance to kanamycin and related aminoglycoside antibiotics (such as neomycin, gentamycin, G418 and hygromycin) conferred by a bacterial aminoglycoside phosphotransferase gene. Kanamycin is particularly widely used with dicotyledonous plants; but many cereals are tolerant to this, and in such cases G418 and hygromycin are often used instead. The herbicide phosphinothricin (also known in modified forms as Basta, Liberty or bialaphos) inhibits glutamine synthetase. Two genes are available as selectable markers conferring resistance to phosphinothricin. These are the *bar* and *pat* genes, which were isolated from resistant microorganisms. They both encode a phosphinothricin acetyltransferase, which inactivates the

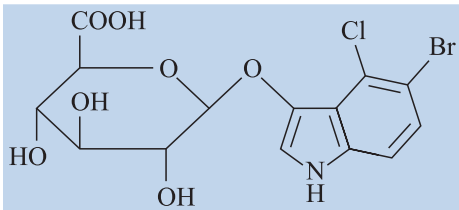


Fig 9.13 Structure of X-gluc (5-bromo-4-chloro-3-indolyl-β-D-glucuronide).

herbicide, and the genes can, therefore, be used as selectable markers when placed under the control of plant promoters. The use of markers conferring resistance to herbicides and antibiotics in transgenic crop plants has caused public concern, on the grounds that these genes may be transferred laterally. Two main approaches have been taken to deal with this. One has been to develop systems for the elimination of marker genes, in many cases by the Cre-*lox* system (see Chapter 7) once stable transformation has been achieved. The other approach has been to use markers that do not require antibiotics for selection. One strategy exploits the fact that many plants are unable to use the sugar mannose. The phosphomannose isomerase gene allows plants to grow on mannose as a carbon source as it converts mannose-6-phosphate to fructose-6-phosphate, which can be metabolized. A xylose isomerase gene has been used in a similar way, allowing growth on xylose as carbon source.

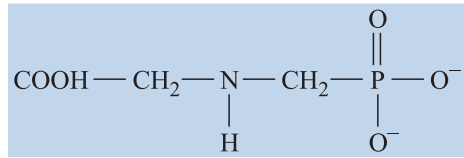
9.6.5 Reporter genes

Genes for the fluorescent proteins, such as GFP and the other fluorescence colours, are widely used as reporter genes in plants. The gene for beta-glucuronidase was used widely before the development of GFP. Beta-glucuronidase hydrolyses an artificial substrate X-Gluc (5-bromo-4-chloro-3-indolyl-β-D-glucuronide; Figure 9.13) to generate a blue pigment (analogous to the hydrolysis of X-gal by beta-galactosidase described in Chapter 3). There is very little, if any, endogenous beta-glucuronidase activity in most plant tissues, but tissue sections can be readily stained for the enzyme with X-Gluc and plant extracts assayed for it. This makes the enzyme particularly useful as a reporter. Other substrates, such as 4-methyl-umbelliferyl-β-D-glucuronide, can also be used. In this case the product is measured fluorometrically. Other reporter genes include those for luciferase, which is assayed by light production, and chloramphenicol acetyltransferase, assayed by its ability to transfer radioactive acetyl groups onto chloramphenicol.

9.6.6 Applications

Methods for the genetic manipulation of plants have, of course, been very useful for understanding plant biology. However, they have had

Fig 9.14 Structure of glyphosate.



a number of practical applications, in the development of crop plants with novel characteristics. We will look at a few particularly interesting examples.

1. **Herbicide resistance/tolerance.** This is usually achieved by causing plants to overexpress the target protein for a herbicide in order to titrate the herbicide out, to produce a protein of modified sequence that is now resistant or to produce an enzyme to detoxify the herbicide. Resistance or tolerance to the phosphinothricin herbicides can be conferred by expression of an acetyl-transferase gene. Another herbicide of great interest is glyphosate (also known as Roundup; Figure 9.14) which inhibits the enzyme 5′-enolpyruvylshikimate-3-phosphate synthase (EPSP synthase), an important enzyme of the shikimic acid pathway for the synthesis of aromatic amino acids. Glyphosate resistance can be engineered by the expression of a mutant protein that is tolerant to glyphosate and a bacterial glyphosate oxidoreductase that degrades it. A number of important crops have been engineered to glyphosate resistance, including maize, oilseed rape (canola), soya and beet.
2. **Virus resistance.** Causing plants to overexpress viral coat proteins often confers resistance to the corresponding virus, perhaps through interfering with the infection process. Tomatoes, peppers, cucumbers and papaya are among the plants that have been engineered to virus resistance in this way. Virus resistance can also be engineered using RNAi methods (see Chapter 7).
3. **Insect resistance.** Many strains of the bacterium *Bacillus thuringiensis* produce proteinaceous insecticidal toxins during sporulation, and these can be used directly as insecticides. Genes for the toxins can be incorporated into plants under appropriate promoters to cause the plants to become insecticidal themselves. Cotton was the first major crop to be engineered to insect resistance in this way.
4. **Control of ripening.** Ripening in tomatoes depends on the production and response to ethylene. This is produced from 1-aminocyclopropane-1-carboxylate (ACC) by the ethylene-forming enzyme. ACC is produced by ACC synthase (Figure 9.15). Inactivation of expression of either of these enzymes by use of antisense RNA generates plants yielding fruit that do not ripen of their own accord (as they are unable to produce ethylene),

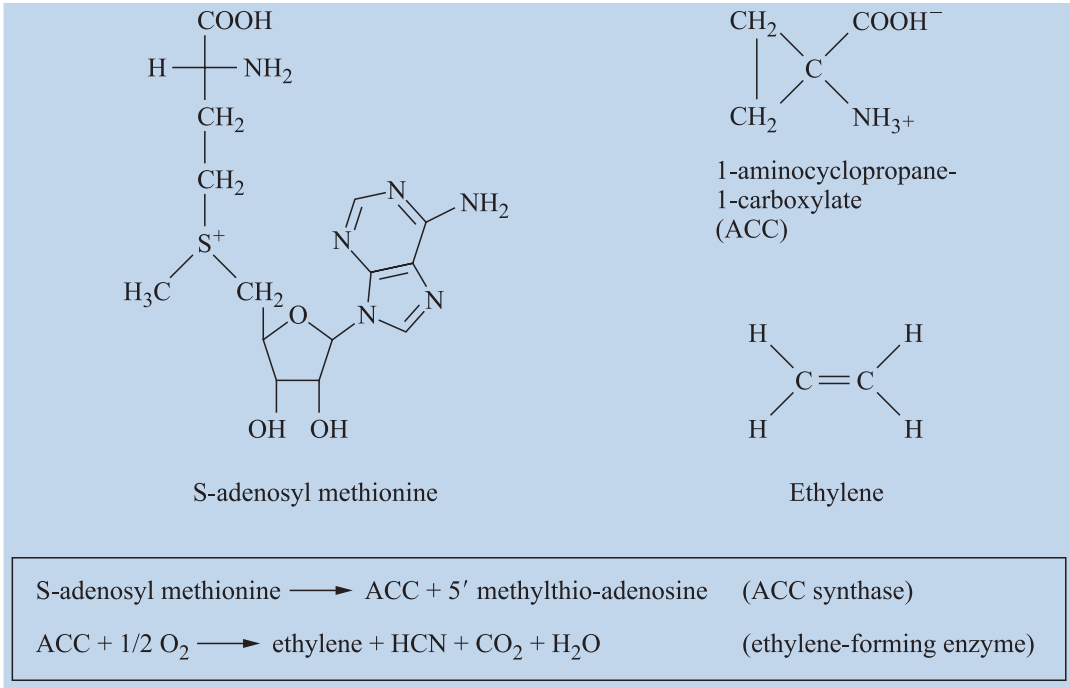


Fig 9.15 Synthesis of ethylene from *S*-adenosyl methionine via 1-aminocyclopropane-1-carboxylate (ACC).

but they will do so if given exogenous ethylene. The 'Flavr Saver™' tomato produced by Calgene contained an antisense gene for polygalacturonase, an enzyme that causes wall softening during the ripening process. The tomato was also found to have improved processing qualities.

5. **Male sterility.** Male-sterile plants (that do not produce viable pollen) are an important tool in plant breeding, to avoid self-pollination. Incorporation of a gene for a bacterial ribonuclease (barnase) attached to a promoter specific for the tapetum (tissue surrounding the pollen sac, which is important in pollen development) leads to failure of pollen formation and male sterility. This forms the basis of terminator technology used by plant breeding companies to stop the propagation by farmers (or in the event of 'escape' into the wild) of transgenic crops.
6. **Other physiological modifications.** A huge range of other physiological modifications has been achieved. These include developing resistance to chilling, drought, salinity and fungal and other pathogens. The nutritional quality of grain can also be enhanced, with important implications for developing countries. Inactivation of genes for enzymes in caffeine biosynthesis has been suggested as a way of generating 'naturally' decaffeinated coffee beans.
7. **'Pharming'.** Plants are attractive as hosts for the expression of pharmaceutical products. They have been used for the expression of vaccines, recombinant antibodies and other pharmaceuticals, such as hormones.

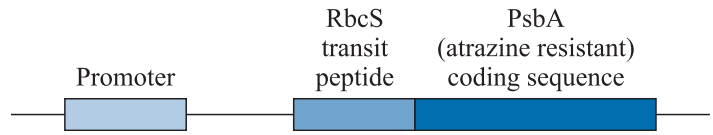


Fig 9.16 Construct used for allotopic expression of *psbA*. Details are given in the text.

9.7 Organelle transformation

The examples we have covered so far have dealt with the modification of the nuclear genome. However, eukaryotes contain one or two additional genomes, located in the chloroplast and the mitochondrion. They encode a small but significant part of the polypeptides of those organelles. The chloroplast genome encodes mainly polypeptides involved in the light reactions of photosynthesis, protein synthesis and a limited number of other functions. The mitochondrial genome encodes mainly polypeptides involved in oxidative phosphorylation. In addition, both organelle genomes encode rRNAs and tRNAs. Mutations in mtDNA frequently result in the cell's inability to carry out oxidative phosphorylation, and we will see later that this has been exploited in mitochondrial transformation work.

Early attempts to manipulate organelle genomes included the use of *Agrobacterium* and direct DNA uptake, but they met with limited success. It was difficult to demonstrate DNA uptake conclusively and reproducibly. For some purposes, a different approach could be taken by fusing to a gene a sequence that encoded an organelle-targeting peptide and then inserting the modified gene into the nucleus, rather than into the organelle. An example of this was the engineering of tolerance in tobacco to the herbicide atrazine. This herbicide acts on one of the polypeptides of the reaction centre of Photosystem II, which is the product of the chloroplast *psbA* gene. A construct was made containing a mutant *psbA* gene (whose product was resistant to atrazine) from the plant *Amaranthus* fused to a sequence encoding the transit peptide of the small subunit of ribulose bis-phosphate carboxylase. (This polypeptide is encoded in the nucleus, synthesized in the cytoplasm and subsequently imported into the chloroplast. The transit peptide is a region at the amino-terminus of the protein, which directs import into the chloroplast and is subsequently removed proteolytically.) This hybrid gene, with a plant nuclear promoter attached (Figure 9.16), was introduced into the tobacco nucleus by *Agrobacterium*-mediated transformation. The PsbA protein was produced and imported into the chloroplast under the direction of the transit peptide, and the plants showed a significant level of atrazine resistance. Similar approaches can be used to target other polypeptides to the chloroplast or the mitochondrion as appropriate, and if the presence of the normal gene in the organelle is not a problem, then inserting modified organelle genes into the nucleus may well be suitable. This is sometimes called allotopic gene expression. For some

species, direct modification of the chloroplast and mitochondrial genomes is possible.

9.7.1 Chloroplasts of algae

The first chloroplast manipulation experiments used *Chlamydomonas* strains mutant in the chloroplast gene *atpB*, encoding the beta subunit of the ATPsynthase, an important multisubunit complex required for the generation of ATP using light energy. The mutant had a partial deletion of this gene, rendering it unable to grow photosynthetically and dependent on a suitable fixed carbon source, such as acetate, in its growth medium. Cells were bombarded with a plasmid containing the cloned wild-type gene, incubated for a while, then transferred to medium lacking a fixed carbon source and illuminated, thereby selecting for the restoration of photosynthetic competence. Photosynthetic colonies were obtained, containing an apparently normal beta subunit in the ATP synthase, and with the mutant region of chloroplast DNA replaced with the wild-type region, as judged by Southern analysis. This replacement was presumably the result of homologous recombination, and all copies of the chloroplast genome contained the wild-type sequence, since restriction fragments corresponding to the mutant could not be detected. The lack of any copies of the mutant allele is remarkable because the chloroplast DNA is present at a high copy number, and it indicates that the wild-type phenotype will be stable in the absence of continued selection.

Other selectable markers (with suitable chloroplast expression sequences attached if necessary) that have been successfully used in chloroplast transformation in *Chlamydomonas* include:

(a) a bacterial *aadA* gene, encoding an aminoglycoside adenyltransferase and conferring resistance to spectinomycin or streptomycin;
(b) *aphA-6*, an aminoglycoside phosphotransferase conferring kanamycin resistance;
(c) restoration of photosynthetic competence in strains with a mutation in a gene for one of the proteins involved in the light reactions of photosynthesis;
(d) resistance to streptomycin, spectinomycin or erythromycin conferred by mutations in the 16S or 23S rRNA genes.

The selectable markers can also be used as the basis for gene disruption. Reporter genes, such as those for green fluorescent protein and firefly luciferase, are available. An advantage with the use of *Chlamydomonas* for transformation (and especially for gene disruption) is that, unlike higher plants, the organism has a single chloroplast; therefore, there is no risk of getting cells that contain a mixture of chloroplasts, some of which have been genetically altered and some of which have not. Such mixed cells are said to be heteroplasmic, rather than homoplasmic. There is still the possibility

of getting mixed populations of genomes within the single chloroplast, and this may occur if disruption of all copies of the gene would result in cell death. Biolistic transformation is the most widely used way of introducing DNA into the *Chlamydomonas* chloroplast, but chloroplast transformation by rapid mixing of protoplasts with DNA and glass beads, or electroporation, has also been reported. Chloroplast transformation systems have been established for other species, including *Euglena* and the red alga *Porphyridium*.

9.7.2 Chloroplasts of higher plants

Transformation of chloroplasts of higher plants was first reported definitively with tobacco, a couple of years after the description of transformation of *Chlamydomonas* chloroplasts. The usual approach is to use a vector that cannot replicate in the chloroplast, but contains a selectable marker and the gene of interest flanked by stretches of chloroplast DNA that are adjacent in the chloroplast genome. The vector is introduced by biolistic transformation. Homologous recombination across the chloroplast DNA sequences allows the integration of the selectable marker and the gene of interest. A few vectors that contain chloroplast origins of replication and can replicate independently of the chromosome have also been developed, and transient expression systems have also been used. The most widely used selectable markers are antibiotic resistance genes attached to appropriate plastid expression signals. They include *aadA*, encoding aminoglycoside adenyltransferase and conferring resistance to spectinomycin and streptomycin, and a variety of aminoglycoside phosphotransferase genes, conferring resistance to neomycin or kanamycin. These markers can be used for gene disruption experiments. There is also a counter-selection system, based on the cytosine deaminase gene. The product of this converts 5-fluorocytosine to the toxic 5-fluorouracil. If the gene has been introduced into the chloroplast, its subsequent loss can be selected by growth in the presence of 5-fluorocytosine. Genes for fluorescent proteins and for beta-glucuronidase are used as reporters.

Chloroplast transformation has been reported for a number of higher plants, as well as tobacco. These include rice, potato, tomato, petunia, cotton and *Arabidopsis*. It has been suggested that incorporation of transgenes into the chloroplast may offer more biological containment than when transgenes are incorporated into the nucleus. This is because in many important crop lines, the chloroplast is not transmitted through the pollen. However, it has been shown that chloroplast DNA sequences can transfer to the nucleus at a surprisingly high frequency and, therefore, evade this containment. As with manipulation of the nuclear genome, concern has also been expressed at the widespread release of antibiotic resistance marker genes when genetically modified crops are planted. This has led to the development of a Cre-lox system for removal of marker genes after stable transformation. In this,

the construct used for transformation includes a marker gene flanked by *loxP* sites. After stable transformation has been established, the transgenic line is crossed to a line that contains a nuclear gene for a chloroplast-targeted Cre recombinase. This catalyses the loss of the marker gene from the chloroplast.

9.7.3 Yeast mitochondria

Transformation of *Saccharomyces cerevisiae* mitochondria can be achieved using biolistic transformation, and was first reported in 1988. This example used yeast cells that were mutant in the nuclear *URA3* gene and in the mitochodrial *COX1* gene, encoding subunit I of cytochrome oxidase. The mitochondrial mutation rendered the cells unable to carry out oxidative phosphorylation and, consequently, unable to grow on glycerol as a carbon source (as cells cannot grow on glycerol by fermentation alone). The cells were bombarded with particles carrying a mixture of DNAs containing wild-type *COX1* and *URA3* and selected for the acquisition of wild-type *URA3* but not, initially, wild-type *COX1*. Colonies that grew in the absence of exogenous uracil were then plated onto medium containing glycerol as the sole carbon source to select for wild-type *COX1* cells, which were indeed obtained. It was argued that the initial selection for *URA3* acquisition would identify those cells that had received DNA and, from these, the cells that had incorporated the mitochondrial DNA could be selected more easily. As with chloroplast transformation, it seemed that the wild-type sequences were integrated by homologous recombination to replace the defective sequences.

It is often more convenient to introduce DNA first into mitochondria of mutant strains of *Saccharomyces cerevisiae* that have had all their endogenous DNA deleted (such strains are designated *rho*^{0}). The DNA that is introduced into the mitochondria becomes reiterated into concatemers, and the resulting individuals are often called *synthetic rho*$^-$, because they resemble naturally occurring *rho*$^-$ strains, which have had most of their mitochondrial DNA deleted and the remainder formed into a concatemer of repeating units. These synthetic *rho*$^-$ strains can subsequently be used to transfer DNA into wild-type strains, by mating. During this process, mitochondria fuse and recombination takes place between mitochondrial genomes. Generating a synthetic *rho*$^-$ intermediate and then transferring the DNA by classical genetics into a recipient strain for analysis of transgene function has a number of advantages. It does not require the construction of a new transformation host strain for each separate experiment, and transformation of rho^0 mitochondria is more efficient than transformation of mitochondria that already contain DNA. A typical strategy is summarized in Figure 9.17. The gene of interest is inserted into a vector that contains a selectable marker, typically the *COX1* gene. The vector also contains a nuclear selectable marker, usually *URA3*, although this can be supplied

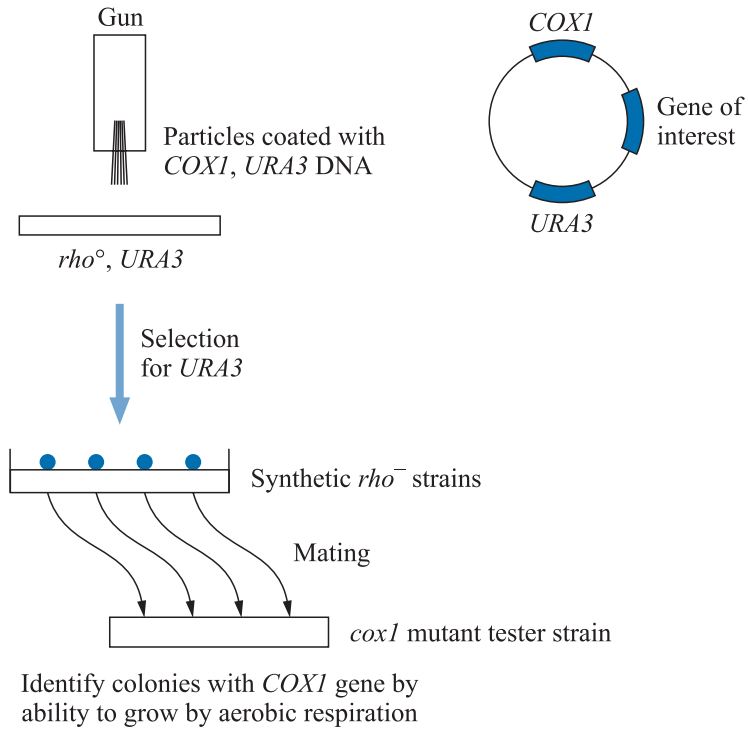


Fig 9.17 Mitochondrial transformation in *Saccharomyces cerevisiae*. Colonies that have acquired *URA3* are tested for the mitochondrial *COX1* gene by mating to a mutant strain lacking functional *cox1*. They also have the gene of interest in the mitochondrion. The gene of interest can then be transferred into other mitochondrial types by mating.

instead on a separate DNA molecule. The DNA is then introduced into a *rho*^{0} host by biolistic transformation, and nuclear transformants selected by growth in the absence of added uracil. This generates a large number of nuclear transformant colonies. Cells from each colony are then picked off and mated with a *cox1* tester strain (i.e. one that has a mutation in its *COX1* gene). If a picked cell gives rise to products in the mating that are able to grow by aerobic respiration, it must carry a wild-type *COX1* gene and, therefore, be a mitochondrial transformant. (Typically, one nuclear transformant in 1000 is also a mitochondrial transformant. Mating with the tester strain can be carried out easily. A lawn of tester strain is poured on a suitable plate and samples of the colonies to be tested are spotted directly on to it. Growth indicates that the colony tested contains the *COX1* gene.) We have, therefore, identified transformed cells that have acquired the gene of interest, and this can be transferred into other strains by conventional crossing. The three stages are therefore:

(a) selection for nuclear transformants using *URA3*;
(b) screening of those to identify mitochondrial transformants;
(c) transfer of the mutant allele into other recipient strains.

The *COX3* gene is also used as a selectable marker. It confers the ability to grow by aerobic respiration on cells with a mutation in the mitochondrial *COX3* gene. Other genes that are not of mitochondrial origin can also be used, after modification to follow the

mitochondrial genetic code. These include the *ARG8*^{m} gene for acetylornithine aminotransferase, an enzyme of arginine biosynthesis. This enzyme is usually located in the mitochondrion, but is the product of a nuclear gene. The *ARG8*^{m} gene is designed for expression in the mitochondrion and will confer the ability to grow in the absence of arginine on cells with a nuclear *arg8* mutation. The *ARG8*^{m} gene is also a convenient reporter gene. Another selectable marker is a modified form of the gene for the RNAse inhibitor Barstar. Use of this marker requires cells carrying a gene for an RNAse, Barnase, carrying a mitochondrial targeting sequence. The Barnase enters the mitochondrion and degrades mitochondrial RNA, blocking mitochondrial function. (If the Barnase is expressed at too high a level, it will also destroy cytosolic RNA.) A Barstar gene designed for expression in the mitochondrion can be used as a selectable marker in such a strain. Incorporation of the Barstar gene into the mitochondrion allows its expression. The resulting Barstar inhibits the Barnase and restores mitochondrial function.

9.7.4 *Chlamydomonas* mitochondria

The first demonstration of transformation of the mitochondrion in *Chlamydomonas reinhardtii* used a respiratory deficient strain, *dum-1*, with a deletion in the mitochondrial gene for *cytochrome b*. This made the strain unable to grow in the dark. Biolistic transformation using microprojectiles carrying partially purified mitochondrial DNA from wild-type cells led to the recovery of cells carrying the wild-type gene for *cytochrome b* and able to grow in the dark. Manipulation of the *Chlamydomonas* mitochondrial genome is not widely used at present, but has great potential for studying mitochondrial biogenesis and function in photosynthetic organisms.

9.8 *Caenorhabditis elegans*

The nematode worm *Caenorhabditis elegans* is of great importance as a model biological system. However, genetic manipulation of the organism is complex. DNA can be injected into germlime tissue or introduced by biolistic transformation, but the DNA typically forms extrachromosomal arrays of multiple copies, which do not behave stably. The incorporation of poison sequences that are toxic if overexpressed can be used to help combat this. Under some circumstances, biolistic transformation can lead to stable chromosomal integration of sequences in single copies. In a few cases, this integration is by homologous recombination, offering the possibility of gene targeting. However, the effects of inactivation of genes are most easily determined using RNAi. Double-stranded RNA can be easily introduced for this purpose simply by feeding it to animals, soaking them in a solution of double-stranded RNA, or by microinjection.

9.9 Insects

9.9.1 Cultured cells and baculoviruses

One of the important applications of insect cell transformation is the use of baculoviruses infecting cultured insect cells as expression vectors. The most commonly used baculovirus is *Autographa californica* nuclear polyhedrosis virus (AcNPV), which was isolated from the alfalfa looper caterpillar and can infect around 30 different Lepidopteran species. The virus has a double-stranded circular genome of 128 kbp. DNA replication takes place about 6 h after infection. By about 10 h after infection, new virus particles are produced by budding from the cell membrane. These are termed extracellular virus particles. Later, occluded virus particles are produced, which accumulate in the nuclei of infected cells as occlusion bodies or polyhedra. The virus particles are held together by a 29 kDa protein called the polyhedrin protein, together with another protein, p10. They serve to protect the virus particles from environmental damage after the death of the host insect and before ingestion by another insect (after which the polyhedra dissolve in the gut allowing the virus to infect surrounding cells).

The polyhedrin protein is produced at extremely high levels (up to 20% of protein synthesis), as is p10, and at this stage in the infection cycle the expression of other genes is low. If the virus is grown in cultured cells, polyhedrin and p10 are non-essential. These facts are exploited in baculovirus expression systems. The principle is that the sequence to be expressed is inserted downstream from the polyhedrin or p10 protein promoter, and the recombinant viral genome is introduced into cultured insect cells, the cell lines most widely used being derived from ovarian tissue of the fall armyworm, *Spodoptera frugiperda*, or the cabbage looper, *Trichoplusia ni*. The polyhedrin or p10 promoter then directs high levels of expression of the inserted sequence. Other approaches have also been taken, e.g. using a nuclear polyhedrosis virus from *Bombyx mori*, then infecting silkworm larvae with the recombinant viruses and recovering protein from the haemolymph. Expression of proteins using nuclear polyhedrosis viruses is advantageous not only because it gives high protein yields, but also because a range of post-translational modifications may be carried out that are carried out much less reliably (if at all) with prokaryotic or yeast systems. These modifications include glycosylation (O- and N-linked), phosphorylation, acylation, proteolytic processing and secretion. One or more of these modifications may be required for the expressed protein to have biological activity.

However, the size of the viral genome presents a problem, as it is too large to be easily manipulated in vitro. A convenient solution is to use a system akin to the co-integrative vector system used with *Agrobacterium tumefaciens*. The baculovirus system requires a small plasmid termed the transfer vector. Figure 9.18 illustrates an example of a typical transfer vector that contains

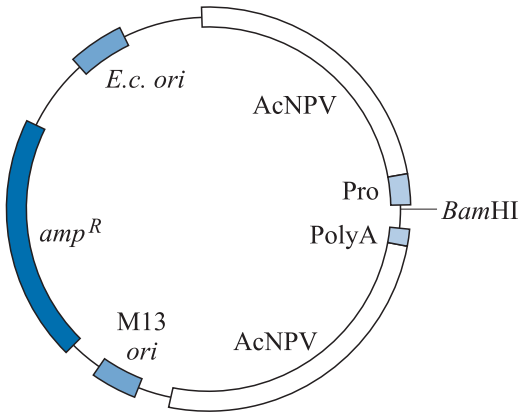


Fig 9.18 Plasmid pBacPAK1 (5.5 kb) from a baculovirus transfer vector system. Sequences can be inserted into the *Bam*HI site between the polyhedrin promoter (Pro) and the polyadenylation site (PolyA). Viral sequences on each side (AcNPV) allow integration into a complete viral genome by recombination. The plasmid can be propagated in *E. coli* using the origin (*E. c. ori*) and ampicillin resistance marker (amp^R), and there is also an origin for single-stranded DNA synthesis if required (M13 *ori*).

the following: sequences to allow propagation in *E. coli*; a viral promoter (in this case the polyhedrin gene promoter); the polyhedrin mRNA polyadenylation signal; and sequences that, in the virus, flanked both ends of the polyhedrin gene. The gene to be expressed is inserted adjacent to the viral promoter (which can be done to generate fusion proteins or intact proteins, depending on the vector used). The construct is then transfected into cultured insect cells, along with complete intact viral genomic DNA, giving rise to infection plaques in the host cells. The transfection can easily be accomplished by incorporation of the DNA into liposomes that fuse with the insect cells, or by co-precipitation of the DNA onto the cells with calcium phosphate. Once the DNA is inside the insect cells, homologous recombination can take place, in which the polyhedrin gene from the intact viral genome is replaced with the region containing the foreign DNA. This recombination occurs between the sequences flanking the polyhedrin gene in the intact viral DNA and the corresponding homologous sequences in the transfer vector. Recombination generates a modified virus containing the foreign DNA. Plaques produced by these modified viruses can be distinguished visually from plaques formed by the wild-type virus. Large quantities of cells can then be grown and infected with the modified virus. Within these infected cells, the viral promoter directs high levels of transcription of the inserted sequence. The presence of the polyadenylation signal ensures that the mRNA produced is polyadenylated. Large amounts of protein are synthesized, which can be harvested. The whole procedure is summarized in Figure 9.19.

Modifications

One potential difficulty with the system just described is the reliance on recombination to incorporate the modified sequences into the viral genome. Various modifications have been made to alleviate this.

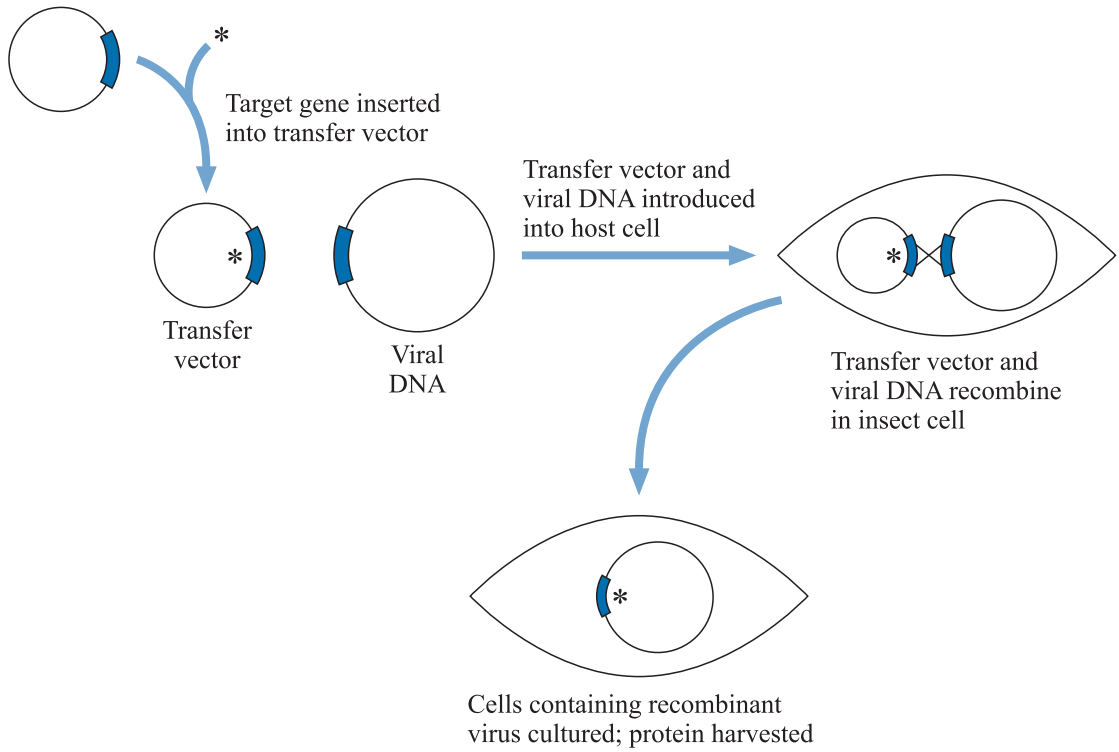


Fig 9.19 Generation of recombinant baculovirus in a host cell. The asterisk represents the gene to be expressed.

Some transfer vectors contain a *lacZ* gene adjacent to the site for insertion of the sequences for expression. The expression in plaques of *lacZ* (detected using X-gal), therefore, indicates that the plaques contain recombinant viruses. This helps in detection of recombinants, but it does not increase their frequency. Modifications to improve the recombination frequency include simply linearizing the DNA, on the grounds that this appears to make it more recombinogenic, and selection schemes whereby regions essential for growth are removed from the viral genome and can only be reconstituted by recombination with the transfer vector. Thus, the only plaques formed should contain the recombinant viruses (as only recombinant viruses have the complete genome needed to infect cells).

A more sophisticated approach allows for the generation of the recombinant virus in *E. coli*. The method uses a virus that has been modified by the addition of a replication origin and selectable marker, so it can be propagated in *E. coli*. In effect, it can function as a shuttle vector, and is sometimes known as a bacmid. Both the bacmid and the transfer (also called the donor) vector contain recognition sites for the recombination system of a transposon such as Tn7 (Figure 9.20). The transfer vector is introduced into an *E. coli* strain that carries the bacmid and a third (helper) plasmid, which encodes the transposase. The transposase catalyses the insertion of sequences from the transfer plasmid into the bacmid, which disrupts a *lacZ′* minigene in the bacmid. *E. coli* colonies containing a recombinant virus can, therefore, be identified by blue-white selection, and the viral DNA recovered

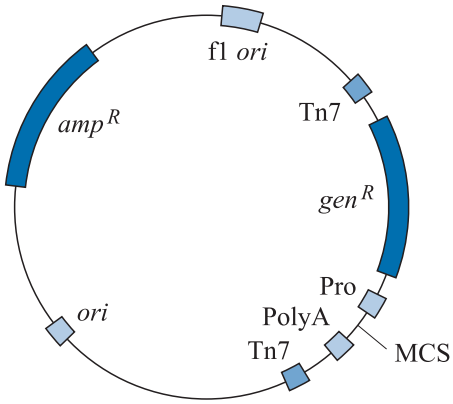


Fig 9.20 The pFastBac™ I transfer plasmid (4.8 kb). Sequences are inserted into the multiple cloning site (MCS) between a polyhedrin promoter (Pro) and an SV40 polyadenylation sequence (PolyA). The region is flanked by Tn7 transposition sequences (Tn7). The plasmid also contains genes for resistance to gentamycin (*gen*^{R}) and ampicillin (*amp*^{R}) and an origin (*ori*) for replication in *E. coli*. There is also an f1 phage origin of replication (f1 *ori*).

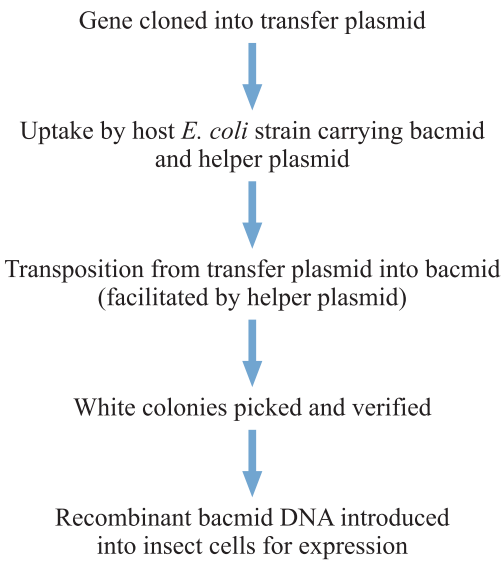


Fig 9.21 Summary of the Bac-to-Bac® cloning system for generating recombinant baculovirus.

and introduced into insect cells. Carrying out the recombination and selection in *E. coli* in this way is faster than using an insect cell host. The procedure is summarized in Figure 9.21.

Enhanced vectors

Various modifications have been made to the basic baculovirus vectors. They allow co-expression of several polypeptides, by the use of several promoters, and secretion of expressed polypeptides under the control of the honeybee melittin secretion signal.

9.9.2 *Drosophila*

Transgenic *Drosophila* can be made by microinjection of DNA into embryos. The embryos are injected at the pre-blastoderm stage, when they contain a layer of nuclei that have not been separated into individual cells. This is called a syncytium, and injecting embryos at this stage means that many nuclei will be accessible to the incoming

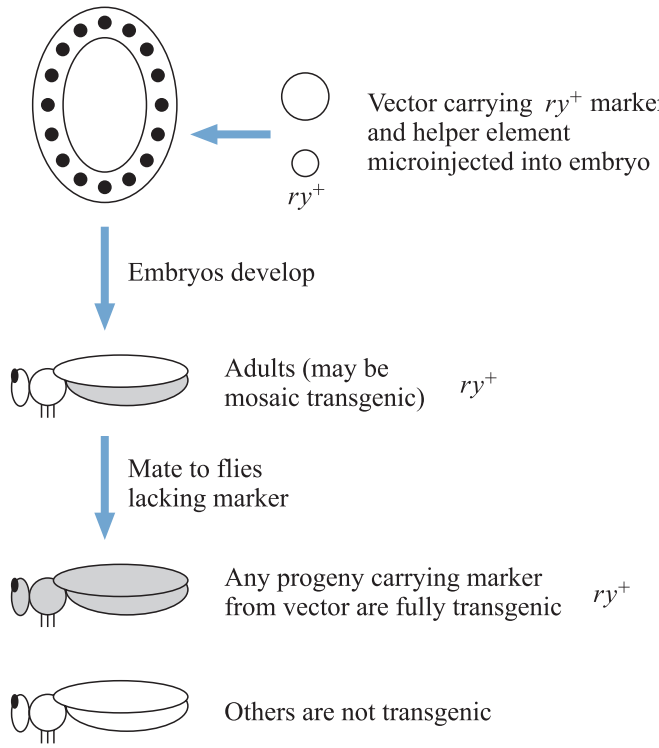


Fig 9.22 Genetic manipulation of *Drosophila*. Mosaic transgenic flies are generated by insertion of an artificial P element into germ-line cells. Progeny derived from transgenic germ cells are fully transgenic. The rosy marker is indicated by ry^+.

DNA. For reasons that will become clear, though, it is only the nuclei that will give rise to germline cells (located at one end of the embryo) that acquire DNA stably. Embryos are then grown to maturity. The mature flies will be mosaic, with some of the germline cells transgenic. Mating these flies produces progeny that are transgenic but no longer mosaic.

In contrast to what happens in many other organisms, injected DNA does not integrate efficiently into the chromosomes of *Drosophila* unaided. Therefore, it is necessary to rely on a 'natural' integration system that depends on transposable elements. For *Drosophila*, the most widely used transposable elements are called P elements, which are the basis of a phenomenon known as hybrid dysgenesis. Under the action of a transposase that is also encoded by the P element, they will insert themselves at random into the DNA of *Drosophila* germline cells. An artificial P element that does not encode transposase, but does have its site of action, can also become integrated if the transposase is provided by a helper P element also present. (This is reminiscent of the binary vector systems used with *Agrobacterium*.) So, in genetic manipulation of *Drosophila* (Figure 9.22), the sequences to be inserted are first incorporated into a modified P element, which needs to contain only a genetic marker and those *cis*-acting sequences needed for transposase action. This element is then micro-injected into pre-blastoderm embryos, together with DNA from a helper P element. This helper element directs the synthesis of the transposase, but lacks the sites for transposase action, so it

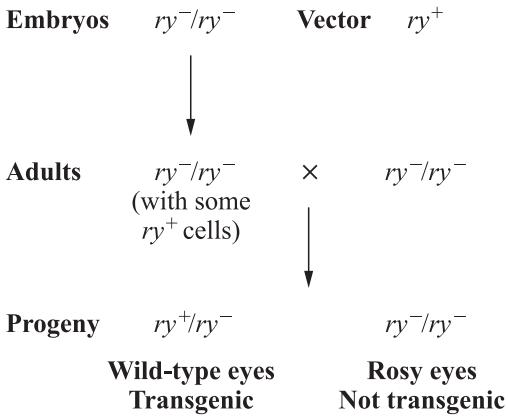


Fig 9.23 Use of the *rosy* gene in manipulation of *Drosophila*.

cannot itself transpose. Microinjection is carried out using a micromanipulator and tiny glass pipettes, although particle bombardment can also be used. Within the syncytial embryo, some of the P element DNAs will find themselves in the region that will ultimately develop into the germline. During that development, the transposase produced from the helper element causes the integration of the modified P element into the germline genome. Transgenic gametes will be produced when the fly is mature, but note that not all the germline cells will necessarily be transgenic, and the rest of the fly will not be — it is a mosaic. To generate a fully transgenic fly, we have to mate the mosaic and obtain progeny flies. Progeny flies that are produced from a non-transgenic gamete from the mosaic will not be transgenic. Progeny formed from a transgenic gamete will be fully transgenic and not mosaic. These flies are identifiable by the presence of the genetic marker on the modified P element. The site of integration of the P element seems to be more or less random, as with *Agrobacterium* T-DNA insertion. Insertion is due to a specific recombinase activity, rather than general homologous recombination.

1. **Markers.** The first marker used for transformation was the wild-type *rosy* gene, ry^+. This encodes xanthine dehydrogenase, which carries out a step in the synthesis of eye pigments. Wild-type flies have brick-red eyes, whereas *rosy* mutants have dark crimson to brown eyes. So, if the host fly embryo which is injected and the fly it is subsequently mated with are both ry^- and the vector carries the wild-type ry^+ gene, transgenic progeny flies can be identified as those with the wild-type eye colour, as shown in Figure 9.23. A disadvantage of *rosy* is that the gene is rather large. Other eye-colour markers are also commonly used, such as *white* and *vermilion*. The *rough* marker is also used, which affects eye morphology. Markers affecting other aspects of *Drosophila* biology are also available, such as the alcohol dehydrogenase adh^+ gene. Wild-type flies can be selected as they are resistant to ethanol supplied in the food, which is toxic to adh^- mutants. A bacterial

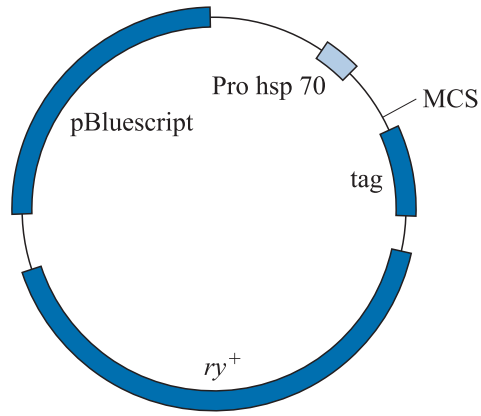


Fig 9.24 *Drosophila* pGS vector series. The vectors are constructed using pBluescript (allowing replication in *E. coli*). The example shown contains an hsp70 promoter (Pro hsp70), a multiple cloning site (MCS) allowing fusion to a tag and the rosy (ry^+) marker.

neomycin phosphotransferase gene attached to a *Drosophila* heat shock promoter is also available. It can be selected by inclusion of the antibiotics G418 or neomycin in the food.

2. **Vectors.** As well as a selectable marker (see above), a vector should contain the appropriate *cis*-acting sites for the transposition process. The promoter from the *Drosophila* heat shock gene *hsp70* is used in many vectors to direct expression of inserted sequences. There are also more complex vectors, which offer epitope tagging and other features (e.g. Figure 9.24).
3. **Helpers and hosts.** Various helper elements are available. The helper elements must encode a functional transposase, and some have been modified to try to increase its production. Helper elements generally have mutations of the *cis*-acting transposition sequences, so that although they can direct the transposition of other sequences, they cannot be incorporated into the transgenic flies themselves. This would lead to genetic instability, as the elements transposed in subsequent generations. The host strain should not contain any integrated P elements, because these might also cause instability. The host must carry suitable mutations for the subsequent detection of incorporated DNA. Tissue-specific expression of inserted sequences can be directed by generating two transgenic lines. One line contains a gene for a transcriptional activator (such as *GAL4*) under the control of a tissue-specific promoter/enhancer. The other line contains the target gene for expression under the control of the recognition sequence for the same transcriptional activator. Crossing the two lines generates progeny with the full system for expression: transcriptional activator, recognition site and target gene. This approach is particularly useful when general overexpression of the target gene is deleterious.

9.9.3 Gene inactivation in *Drosophila*

Drosophila has been an important subject for large-scale studies of the effects of systematically silencing individual genes. Because

P elements integrate at random into the genome, it is not feasible to use them directly for targeted gene disruption. However, the ability of P elements to insert into the genome has allowed the establishment of a large range of lines with genes randomly inactivated by transposon insertion. The recombinogenic activity of linear extrachromosomal DNA has also been exploited to develop recombination-mediated strategies for gene targeting. The most widely used method for gene inactivation is RNAi. Double-stranded RNA can simply be microinjected into embryos (or other developmental stages) for transient gene silencing. For long-term, or heritable, silencing, a transgene is constructed that will direct the synthesis of an inverted repeated copy of the target gene. The transgene is introduced into the target organism and the RNA transcribed from it forms a hairpin that leads to silencing.

9.9.4 Other insects

Similar transposon-based systems using transformation with microinjection have been developed for other insects. Although P element transposition is less efficient in many insects than in *Drosophila*, other transposons (especially the Hermes, mariner and piggyBac elements) have been used. Rather than relying on isolating eye-colour mutants in other species, and using the wild-type genes as selectable markers, most studies have used an artificial gene comprising a coding region for a modified GFP under the control of a synthetic promoter. This promoter comprises three copies of the binding site for a transcription factor, Pax-6, adjacent to a TATA sequence, and is referred to as the 3xP3 promoter. It functions in a very wide range of insects. GFP expression can be detected in the eyes of transgenic insects.

9.10 Mammals

The other animal group we will look at in detail is mammals, although many of the techniques outlined here can be applied to other vertebrates, such as amphibians and fish. There are three kinds of genetic manipulation that we might need to carry out. They involve cultured cells, restricted areas of intact organisms or whole individuals.

9.10.1 Cultured cells

Selectable markers

The first transformations in mammalian systems were achieved with cultured cells from humans suffering from the Lesch–Nyhan syndrome, which is caused by a deficiency of the enzyme hypoxanthine-guanine phosphoribosyl transferase (HGPRT). Cells deficient in HGPRT die in a selective medium containing hypoxanthine, aminopterin, and thymidine (called HAT medium; Figure 9.25). This is because aminopterin blocks the pathway for

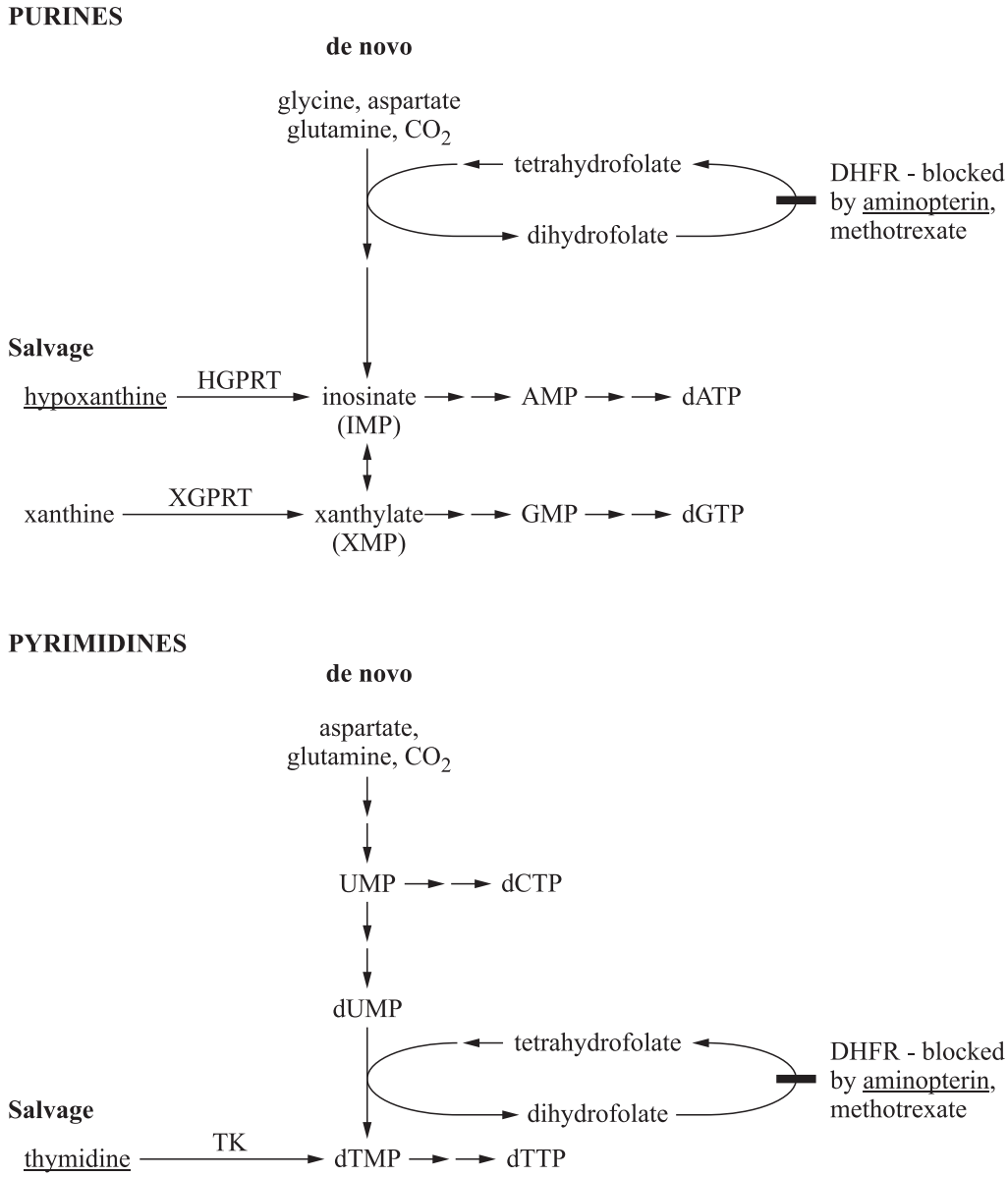


Fig 9.25 HAT selection. Compounds present in HAT medium are underlined. Aminopterin and methotrexate block the *de novo* synthesis of purine and pyrimidine nucleotides; synthesis from hypoxanthine and from thymidine require functional HGPRT and TK enzymes respectively.

endogenous synthesis of the purines needed for synthesis of nucleic acids. An alternative pathway allows synthesis of purines from hypoxanthine, but this requires the enzyme HGPRT, and so cannot operate in $HGPRT^-$ cells. So $HGPRT^-$ cells are unable to make nucleic acids in HAT medium and will die; wild-type cells are able to make nucleic acids and will live. Precipitation of wild-type DNA onto the $HGPRT^-$ cells in the presence of calcium phosphate led to DNA uptake and stable transformation. Similar results could be obtained with cells deficient in thymidine kinase (TK). They are also killed in HAT medium. Pyrimidine synthesis is also blocked by aminopterin, and utilization of the thymidine supplied in the HAT medium

(Figure 9.25) requires a functional TK. Transformation of TK^- cells with DNA containing a functional TK gene gave rise to cells that were able to survive in HAT medium.

So, suitable treatment of cultured cells can lead to uptake and expression of exogenous DNA. In addition to HGPRT and TK genes, commonly used selectable markers for mammalian cell transformation include those listed below. Note that selectable markers derived from prokaryotes have been modified to allow expression in mammalian cells.

(a) Resistance to aminoglycoside antibiotics, conferred by a bacterial aminoglycoside phosphotransferase (also known as neomycin phosphotransferase) gene. The aminoglycoside G418 is widely used for selection.
(b) Resistance to the protein synthesis inhibitor hygromycin, conferred by a bacterial *hph* gene encoding hygromycin phosphotransferase.
(c) Resistance to the protein synthesis inhibitor puromycin, conferred by a *Streptomyces* gene encoding puromycin-*N*-acetyltransferase.
(d) Resistance to the DNA damaging agent bleomycin (zeocin), by expression of a binding protein.
(e) Resistance to methotrexate, which inhibits the enzyme dihydrofolate reductase (DHFR). This enzyme is involved in metabolism of one-carbon units and is required for nucleoside biosynthesis. Resistance can be conferred (in the absence of exogenous nucleosides) by expression of high levels of a sensitive enzyme or by expression of a resistant enzyme.
(f) Resistance to toxic adenosine nucleosides such as 9-β-D-xylofuranosyl adenine. This is conferred by overexpression of adenosine deaminase (ADA) in cells that lack it, or overexpression (in the presence of higher concentrations of 9-β-D-xylofuranosyl adenine) in cells that already contain some ADA.
(g) Resistance to difluoromethylornithine, conferred by high levels of expression of ornithine decarboxylase, the enzyme it inhibits.
(h) A bacterial xanthine-guanine phosphoribosyl transferase (XGPRT) gene. Expression of XGPRT allows $HGPRT^-$ cells (or wild-type cells in the presence of aminopterin) to utilize exogenous xanthine for purine nucleotide biosynthesis.
(i) Histidinol dehydrogenase, which allows cells to synthesize histidine from exogenous histidinol, and protects against toxic effects of histidinol.
(j) Overexpressing glutamine synthetase genes, which allow cells to continue to make glutamine from exogenous glutamate and ammonia in the presence of the glutamine synthetase inhibitor methionine sulphoximine.
(k) Resistance to the peptidyl nucleoside analogue blasticidin, an inhibitor of protein synthesis, conferred by genes that inactivate the antibiotic.
(l) Cytosine deaminase.

As in some other systems, co-transformation is often observed, where cells that have been selected for acquisition of one marker are frequently found to have acquired an unlinked and unselected marker that was supplied at the same time. This means that a selected marker does not necessarily have to be on the same piece of DNA as other transforming sequences. Selection for a particular marker gene can sometimes result in amplification of the number of copies of it in the target cells, which may be useful, although the amplification is not always stable in the absence of continued selection. Counter-selection, for the absence of particular markers, is also possible. The most commonly used counterselectable markers are:

(a) TK from viruses such as herpes simplex virus. This enzyme will metabolize certain nucleoside analogues, such as gancyclovir and bromodeoxyuridine, to form products that damage DNA or block its synthesis. Although many mammalian cells contain TK, the mammalian enzyme uses the analogues much less efficiently than does the viral enzyme. Cells lacking the viral enzyme, therefore, are resistant to the analogues and are able to grow even if they contain the endogenous mammalian enzyme.

(b) Cytosine deaminase (CDA). This enzyme converts 5-fluorocytosine to the toxic 5-fluorouracil. Cells lacking the gene for CDA, therefore, are able to grow in the presence of 5-fluorocytosine. (Positive selection for the presence of CDA is also possible, by growth in the presence of *N*-(phosphonoacetyl)-L-aspartate, which blocks *de novo* pyrimidine biosynthesis by inhibition of aspartate transcarbamoylase, and cytosine, which can provide pyrimidines if CDA is present.)

(c) Diphtheria toxin, which blocks protein synthesis by ADP-ribosylation of elongation factor 2.

Hosts and transformation methods

A wide range of host cell lines is available. Commonly used human cell lines include HeLa, 293T (derived from kidney), and Jurkat (lymphoblasts from a leukaemia patient) cells. Several other tumour cell lines are also quite widely used. Cultured cell lines from other species include COS (derived from kidney cells from African green monkeys), CHO (epithelial-like cells from Chinese hamster ovary), and 3T3 (fibroblasts from mouse) cells. Embryonic stem cells are particularly important for the generation of transgenic whole organisms (see below).

The most commonly used methods for introduction of DNA are calcium phosphate co-precipitation, treatment with DEAE-dextran, liposome packaging and electroporation. The last method (and in some cases DEAE-dextran treatment) is particularly useful for cells growing in suspension. The other methods are best suited for cells growing on a surface. Calcium phosphate co-precipitation is easily achieved by mixing a solution containing the DNA with solutions containing phosphate ions and calcium ions. This results in the

formation of a very fine precipitate of calcium phosphate, particles of which are probably taken up by the cells by endocytosis, with the DNA included passively in the process. Further treatment with glycerol or dimethyl sulphoxide may improve the efficiency of uptake. DEAE-dextran is a modified polysaccharide that is positively charged. This forms a complex with the negatively charged DNA molecules and the complex is taken into the cell by endocytosis. Liposome methods generally involve mixing the DNA with one or more lipids, at least one of which is positively charged. This forms DNA-containing vesicles that fuse with the cell membrane. A wide range of transformation reagents, most of which are based on cationic lipids, are available commercially. Viral packaging, *E. coli* spheroplasts, red blood cell ghosts (cells whose contents have been removed and replaced, by swelling and shrinking in solutions of suitable osmotic strength) and microprojectile bombardment have also been used to introduce DNA into mammalian cells.

Vectors

There are very many vectors available. Some will replicate effectively in cells, allowing extrachromosomal (episomal) maintenance. Others will not and must, therefore, be used in transient expression systems unless they can integrate into the chromosome. There is a range of expression vectors, although they will of course be unnecessary if one wishes simply to test the activity of particular promoter constructs in transformed cells. Overexpression of cloned sequences, on the other hand, can be used for many purposes, including tests of biological function and library screening. Selection for complementation in cells from lines deficient in a particular function allows the identification of cells containing the equivalent cloned gene or cDNA. This has been used, for example, to recover cDNAs for genes involved in DNA repair by complementation of repair-deficient cell lines. If an overexpressed protein is expressed at the cell surface, then it may be possible to select appropriate cells using some kind of panning affinity technique using antibodies or a ligand (see Chapter 6). Clearly, the details of the selection methods used will depend on the particular function of the protein in question and the nature of the cells expressing it. There are also vectors designed for a range of specific purposes, such as two-hybrid screening, targeting of expressed proteins to specific subcellular locations, epitope tagging and reporter systems.

Many vectors are based on the virus SV40. This can replicate successfully in certain cell types, such as those derived from the African green monkey, which are said to be permissive. Additional virus particles can be produced which will go on and infect other cells, although eventually lysis will take place. In non-permissive cells, virus replication does not take place, although the viral DNA can be expressed. If selection for the incoming sequences is imposed, then the only way they can be maintained in non-permissive cells is by integration.

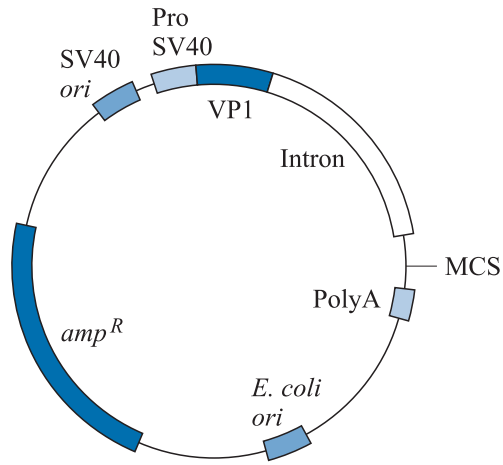


Fig 9.26 Plasmid pSVL (4.9 kb). The vector contains the SV40 late promoter (Pro SV40) and part of the region encoding a viral polypeptide (VPI), with an intron and polyadenylation site (PolyA) associated with VPI. There are SV40 and prokaryotic origins, a multiple cloning site and a prokaryotic selectable marker (*amp*^{R}).

SV40 normally produces two sets of transcripts, denoted early and late. They are initiated divergently from near the origin of replication, and production of the early transcripts is stimulated by two 72-base-pair enhancer sequences. Both sets of transcripts are spliced and polyadenylated. It seems that splicing of transcripts is necessary for efficient expression. Transcripts that have not been through a splicing process are not expressed efficiently, even if introns have been removed from the DNA. So, if vectors are based (as in many examples) upon insertion of sequences into the regions covered by the early or late transcripts, it is important not to disrupt the splicing sites, even if the inserted sequences do not themselves contain introns. An example of such a vector is pSVL, shown in Figure 9.26. This shuttle vector contains pBR322 sequences (from which a 'poison' sequence causing reduced replication efficiency has been deleted) that include the ampicillin resistance gene and the origin of replication. SV40 sequences include: the origin of replication; the late promoter; a region, transcribed from that promoter, that includes intron material; and the polyadenylation site. DNA is inserted after the intron, allowing efficient splicing and polyadenylation. Replication would be possible in permissive cells, giving a high copy number, were it not for the fact that the vector lacks the coding sequence for the T protein (a product of early viral transcription), which is also required for replication. So, in the absence of T, maintenance of the plasmid is possible only by integration. Replication in permissive cells and maintenance extrachromosomally is possible if they are co-infected with a helper virus, or if the cells produce the T protein themselves. This is the case with COS cells (described earlier). These cells will support a high copy number (and, in principle, high expression levels) of SV40-based vectors, although in some cases the copy number is not maintained indefinitely. Control of copy number is possible using a variant of COS cells that produce a temperature-sensitive T protein. Complementation to produce more virus particles (rather than just replication of viral

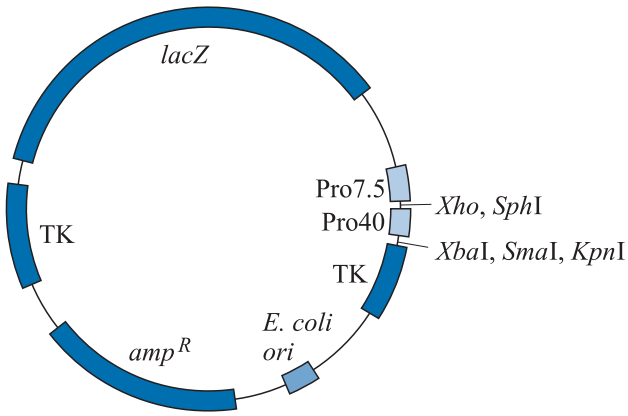


Fig 9.27 Plasmid pAbT4586 (7.8 kb), containing two vaccinia virus promoters (Pro 7.5 and Pro 40). The viral thymidine kinase (TK) sequences allow integration into a viral genome by recombination.

DNA) would require, in addition, the provision of the late viral VP1, VP2 and VP3 proteins, which can be provided by a helper virus. Packaging also places constraints on the size of insert DNA that can be accommodated.

Another widely used virus is the vaccinia virus, a member of the pox virus family. It has a large genome, nearly 200 kb, but can accept at least 25 kb of extra foreign DNA. The virus encodes its own RNA polymerase, which is functional in the cytoplasm of infected cells, where the virus also replicates. Because of the size of the viral genome, a commonly used approach is to generate recombinants in vivo in infected cells by recombination between a suitable plasmid, into which the sequences of interest have been inserted, and a parental virus strain (as with the baculoviruses). Plasmid is transfected into cells carrying the virus and recombination integrates the plasmid into the virus. Techniques are available for the selection of recombinant viruses, e.g. by conferring the ability to replicate in suitable cell lines and/or in the presence of selective agents. Recombinant vaccinia viruses can be used as live vaccines, as well as for the expression of genes in cultured cell lines.

Figure 9.27 gives an example of a plasmid vector, pAbT4586, for use in the vaccinia system. This contains two viral promoters (Pro7.5K and Pro40K) with cloning sites adjacent to them, a modified *lacZ* marker gene and a non-functional viral TK sequence. The plasmid integrates into the viral genome by recombination across the plasmid and viral TK sequences, rendering the virus TK^- because the endogenous viral TK gene is disrupted. TK^- host cells are used, which can grow in the presence of 5-bromodeoxyuridine. Host cells carrying non-recombinant virus will be TK^+, and will be killed by 5-bromodeoxyuridine. Cells carrying a recombinant virus will still be TK^- and will survive in the presence of 5-bromodeoxyuridine. Virus plaques should be blue in the presence of X-gal, because of the beta-galactosidase gene.

Retroviruses are also useful for genetic manipulation of cells in culture. They have the advantages that they will infect susceptible cells with very high efficiency and bring about stable integration into

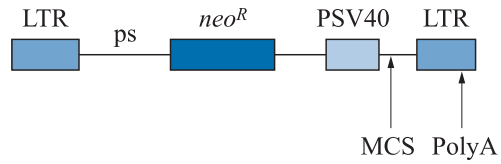


Fig 9.28 Part of the retrovirus-based vector LNSX. (For simplicity, only the retroviral part of the vector is shown.) LTR = long terminal repeat; ps = packaging signal; *neo*^{R} is a neomycin resistance gene; PSV40 is a promoter from SV40.

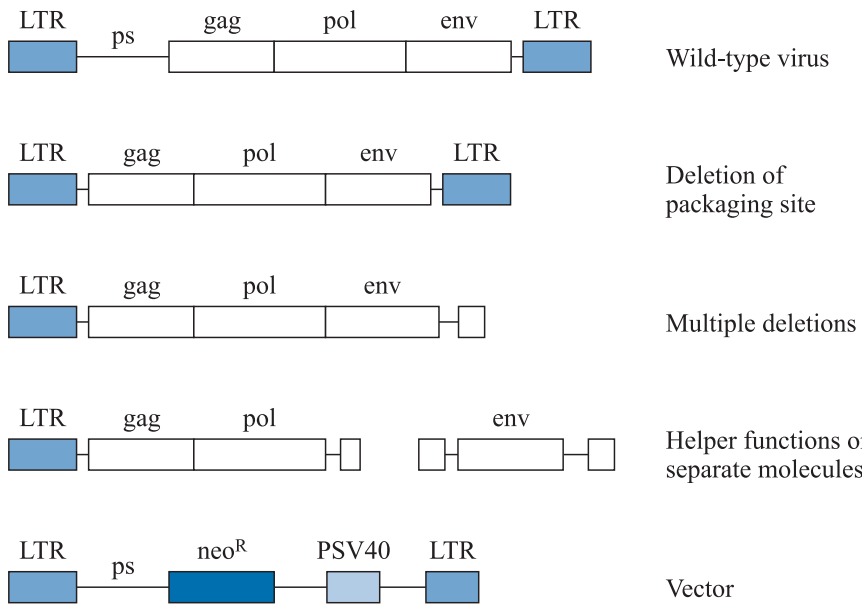


Fig 9.29 Strategies for avoiding packaging of helper retrovirus. The top and bottom diagrams show wild-type virus and a typical vector molecule (see Figure 9.28) respectively. The middle diagrams show three helper retrovirus systems that can provide all the functions for vector packaging, but cannot themselves be packaged. *Gag*, *pol* and *env* are retroviral genes; ps = packaging signal.

the target cell genome. An example, LNSX, is shown in Figure 9.28. This retrovirus contains a neomycin resistance gene as a selectable marker and a promoter (in this case the SV40 early promoter) adjacent to a cloning site. It is usually combined with a prokaryotic plasmid to create a shuttle vector. It does not encode viral proteins, so it cannot itself integrate directly into a host genome. After any construction work in a prokaryotic host has been completed, the retroviral vector is transfected as naked DNA into mammalian cells (*packaging* cells) infected with a helper retrovirus. In these cells, RNA transcripts are produced from the incoming retroviral vector DNA, and these are then packaged into retroviral particles containing the proteins (encoded by the helper retrovirus), such as reverse transcriptase, needed for subsequent infection of target cells. The resulting packaged viruses can then be used to infect other target cells efficiently, resulting in reverse transcription of the retroviral RNA and stable integration of DNA into the recipient cell genome. It is important to stop the helper

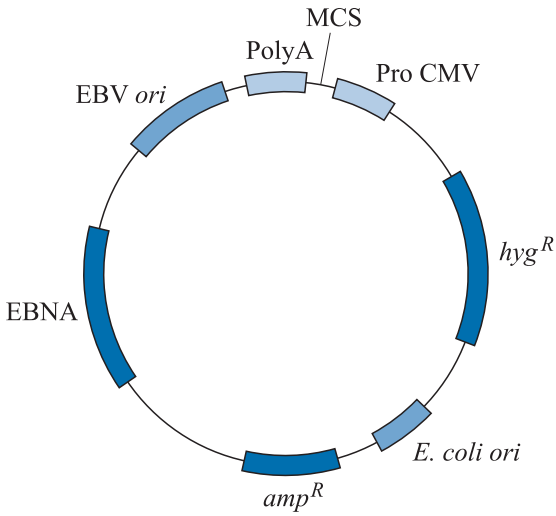


Fig 9.30 Plasmid pCEP4 (10.4 kb), based on Epstein–Barr virus. The plasmid contains an origin of replication from the virus (EBV*ori*) that, together with the nuclear antigen (EBNA), allows maintenance as a plasmid. The plasmid can be selected by a hygromycin resistance (*hyg*^{R}) gene expressed using a thymidine kinase promoter and polyadenylation signal, and sequences in the multiple cloning site can be expressed with the cytomegalovirus promoter (Pro CMV) and polyadenylation signal (PolyA). The plasmid can replicate as a shuttle vector in *E. coli*.

retrovirus from being packaged as well. This can be done by three methods (Figure 9.29): deletion of the packaging signal in the helper genome, deletion of multiple parts of the helper genome, or division of the helper genome into several pieces.

Other viruses used as a basis for vector construction include adenovirus, polyoma virus, which is closely related to SV40, both being members of the papovavirus family, Epstein–Barr virus and bovine papilloma virus (e.g. Figure 9.30).

In the last few years there have been significant advances in the development of MACs (or human artificial chromosomes), capable of carrying large stretches of DNA (in excess of 100 kb) and which can be maintained in the long term without the requirement for selection. A gene that is located within a large stretch of DNA may be more effectively regulated than a gene in an anomalous genomic environment, and this may be important for medical applications. A number of different approaches have been taken to the construction of human artificial chromosomes. In general, they involve either the 'trimming down' of an existing chromosome, or the assembly of a new chromosome from individual components. The most important of these components is the centromere, and in human artificial chromosomes this is usually constructed from tandemly repeated copies of a 171-bp alpha-satellite DNA sequence, found at the centromeres of primate chromosomes and containing recognition sequences for centromere-binding proteins. The generation of these artificial chromosomes is technically very difficult, and this is in part due to the need to manipulate very long stretches of DNA in vitro.

Promoters

CONSTITUTIVE PROMOTERS

Promoters used in any of the vectors described can be divided conveniently into a number of groups: those that are constitutive and those that are induced in response to various signals.

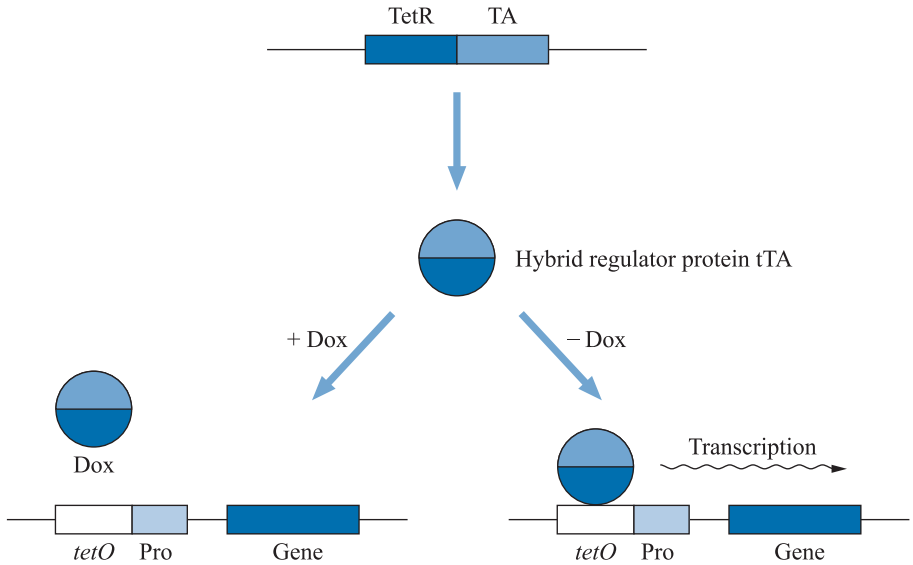


Fig 9.31 The Tet-off system. The hybrid regulator protein tTA comprises the tetracycline repressor domain (TetR) and a transcriptional activator domain from herpes simplex virus, TA. In the presence of doxycycline, the hybrid regulator is unable to bind at *tetO* and transcription from the promoter Pro is inactive. In the absence of doxycycline, the regulator binds and transcription is activated.

Constitutive promoters are frequently derived from viruses. They include the SV40 promoters, the Rous sarcoma virus promoter, the adenovirus major late promoter and the cytomegalovirus immediate early promoter. Hybrid promoters exist, containing elements from, for example, cytomegalovirus and human immunodeficency virus. Enhancers are often included, too, and frequently used ones are derived from SV40, Rous sarcoma virus and cytomegalovirus. However, some enhancers confer a significant degree of tissue specificity and may be unsuitable for general constitutive expression vectors. A number of vectors contain constitutive promoters derived from mammalian genes, such as elongation factor 1 alpha or ubiquitin, rather than viral ones.

TETRACYCLINE-REGULATED PROMOTERS

A very effective control system uses a mechanism for the control of gene expression in response to tetracycline (or analogues) that is derived from the bacterial transposon Tn10. In the naturally occurring bacterial system, a tetracycline-binding repressor protein, TetR, binds in the absence of tetracycline to a control sequence, *tetO*, and inactivates transcription. The method as it has been modified for artificial control of gene expression in mammalian cells (Figure 9.31) has two components. One is the TetR protein fused to part of a transcription-activator protein from herpes simplex virus. This is known as the transactivator protein, tTA. The second component is a hybrid promoter containing the prokaryotic *tetO* control sequence fused to a mammalian promoter (originally derived from cytomegalovirus) with low basal activity. In the absence of tetracycline (or an analogue, doxycycline) the tTA protein binds to the *tetO* control sequence and, because of the herpes simplex virus activation domain,

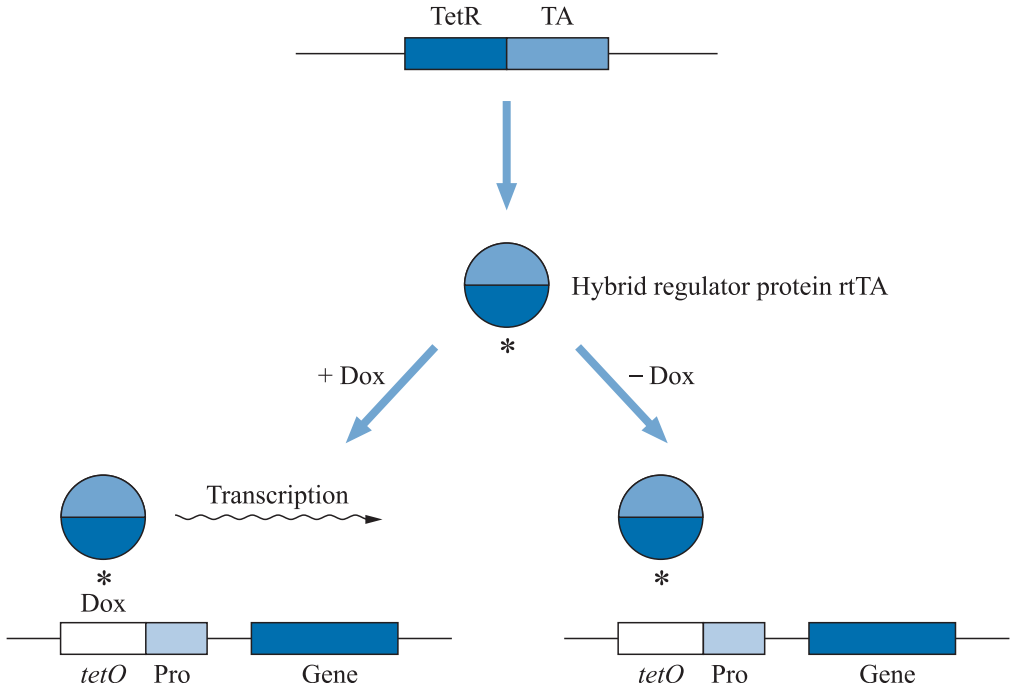


Fig 9.32 The Tet-on system. Abbreviations are as for Figure 9.31. In the Tet-on system, the hybrid regulator binds in the presence of doxycycline, activating transcription.

activates transcription from the mammalian promoter. (At first sight, this may seem confusing, because the original function of TetR is as a repressor! Here, it has been turned into an activator by fusion with the herpes simplex virus transcription activator protein.) The system, therefore, requires a host cell that has been engineered (i) to express the tTA protein and (ii) contains the gene to be expressed under the control of the hybrid promoter. Expression of the gene is inactivated by addition of the tetracycline analogue doxycycline. This system is particularly useful, as the control of expression is very tight – there is very little expression in the presence of doxycycline, so it is very good for expressing proteins that would otherwise be toxic. This is sometimes referred to as the Tet-off system.

The system has been modified by alteration of a few residues of the TetR part of the hybrid control protein (indicated by the asterisk in Figure 9.32). This modified protein is referred to as rtTA, as it has the reverse DNA-binding properties of tTA. It binds target DNA only in the *presence* of doxycycline, and not in its absence. The result of this is that a gene under the control of the hybrid promoter is usually inactive. If doxycycline is added, then it becomes active. This is sometimes referred to as the Tet-on system (Figure 9.32).

GLUCORTICOID-INDUCIBLE PROMOTERS

A widely used inducible promoter is the glucorticoid-responsive element from the long terminal repeat (LTR) of the mouse mammary tumour virus (MMTV). The LTR contains a glucorticoid-responsive element, to which a hormone-receptor complex can bind, activating transcription from the promoter. An example of a vector based on

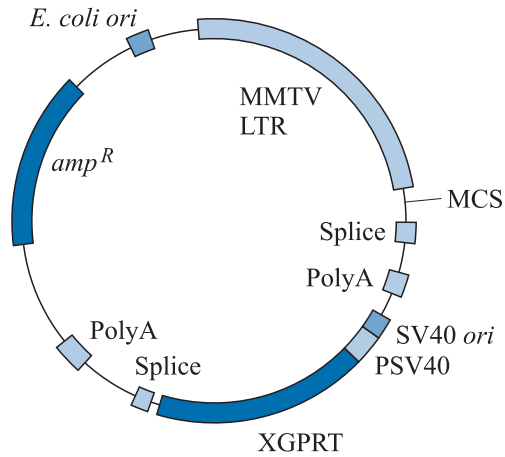


Fig 9.33 Plasmid pMSG (7.6 kb). This contains a mouse mammary tumour virus LTR sequence (MMTV LTR), including a glucocorticoid-inducible promoter. There are also SV40 splice and polyadenylation sites (Splice, PolyA). An SV40 origin—promoter region (SV40 *ori* and PSV40) drives expression of a bacterial XGPRT gene for use as a selectable marker in mammalian cells. There are also splice and polyadenylation sites to aid the XGPRT expression, and a prokaryotic origin and marker for propagation in *E. coli*.

this is pMSG (Figure 9.33), which includes elements from SV40 and pBR322, and a selectable marker (the *E. coli* xanthine-guanine phosphoribosyl transferase gene under the control of the SV40 early promoter). Induction of expression requires treatment with a suitable hormone (or analogue), such as dexamethasone, and the presence of the receptor in the cells. This can be provided either by using a cell type that is already hormone responsive, or by artificial expression in the cells of a receptor protein gene. This approach can be modified to improve expression further by incorporation of several copies of the glucocorticoid response element in the vector.

METAL-INDUCIBLE PROMOTERS

Another important group of promoters comes from the genes for metallothioneins. These are a group of proteins rich in cysteine residues, which chelate heavy metals to reduce their toxicity. The promoters can be induced by treatment of cells with heavy metal ions, such as zinc and cadmium. Inducibility depends on the presence of binding site for a control protein, and this can be incorporated into other promoters too. Use of metal-inducible promoters has two disadvantages: the heavy metal ions used may be damaging to the cells involved; and the control is not particularly tight, so that there may be significant levels of expression in the absence of the metal ions.

HEAT-SHOCK INDUCIBLE PROMOTERS

When cells are exposed to a heat shock, a specific set of genes is activated. This is mediated by the binding of a transcription factor to an element upstream from the promoters. This element can be used either with its normal promoter or in association with others to give controlled expression.

Other considerations

The amount of expression of DNA inserted into a vector does not depend only on the promoter used. If sequences are integrated, then

their chromosomal location is important. Flanking the inserted gene by A elements (or scaffold attachment regions), which are sites for attachment of the chromatin to the chromosomal scaffold, may help to reduce the position effects resulting from chromosomal location. A similar result may be obtained by the use of very long stretches of transforming DNA carried by artificial chromosomes. The presence of introns in the sequences to be expressed is also important, as are other factors such as poyadenylation of the transcript and efficient translation initiation.

9.10.2 Restricted areas of intact organisms

Introduction of DNA into a small region of tissue from an intact organism may be useful, e.g. in studies involving expression or in gene therapy (the introduction of DNA into an individual to treat a genetically determined disorder). For ethical reasons, it is generally considered that genetic modification of humans should not involve the germ line.

The principles for vector construction already outlined apply, although we are more likely to be dealing with a transient expression system. There are many ways of introducing DNA into target tissue, including microprojectile bombardment and direct injection. The latter is particularly suitable for some tissues, such as muscle, which forms a syncytium. The DNA may be complexed into liposomes with conventional transfection agents, and inhalation of liposomes has been used as a means of introducing DNA into the airways in an attempt to treat conditions such as cystic fibrosis. DNA introduction by viral infection is also widely used. For human gene therapy, the most commonly used viruses are retroviruses, adenovirus, adeno-associated virus and herpes simplex virus. There are concerns over the safety of viral vectors in humans, however.

9.10.3 Whole organisms

Generation of transgenic organisms

There are three main methods for generating whole organisms that are transgenic, summarized in Figure 9.34. One is to treat cleavage-stage embryos in vitro with recombinant retroviruses and then return them to the uterus of a foster mother. A more widely used approach is to isolate one-cell embryos (at which stage they are in the oviduct) and micro-inject DNA directly into the pronucleus. The embryos are then returned to a foster mother's oviduct, where development takes place. This approach usually results in the integration of DNA into the genome at random sites (which may also interfere with the function of endogenous genes), and sometimes multiple sites. A third approach exploits the ability to grow and manipulate mouse embryonic stem cells (ES cells) from the inner cell mass of a developing blastocyst. These pluripotent cells can be genetically modified in vitro and then reinjected into the blastocoel of a developing embryo. The injected embryo is then transferred to

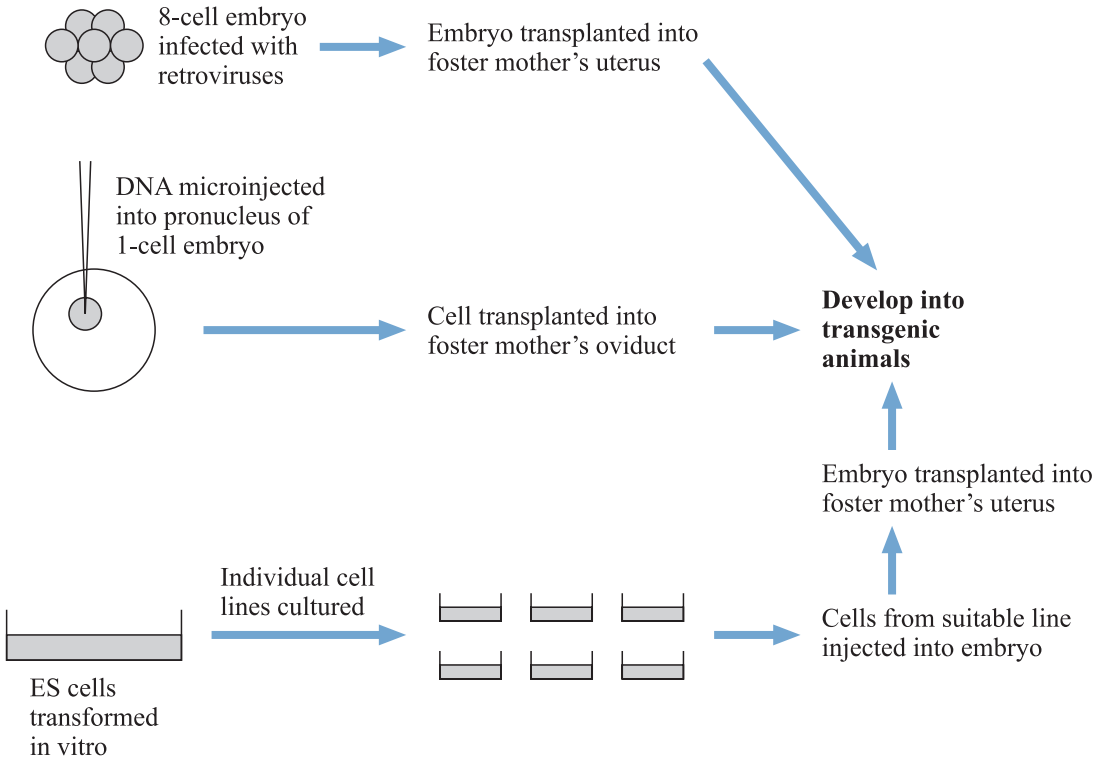


Fig 9.34 Three strategies for generating transgenic animals. See text for details.

a foster mother's uterus. The ES cells colonize the embryo during development, but the embryo does not become entirely composed of ES-derived cells, and the resulting mice will therefore be chimeric. Some ES-derived cells will be incorporated into the germ line. Matings with such mice can generate progeny that are transgenic and not chimeric, if a progeny animal is produced from a transgenic gamete. Producing transgenic mice in this way is analogous to the generation of transgenic *Drosophila* described above. Use of suitable genetic markers, such as coat colour, allows chimeric individuals to be identified easily. Very large stretches of DNA can be acquired, including artificial chromosomes.

The use of the three approaches, especially the last two, for the random integration of incoming DNA is relatively straightforward. To bring about expression of incoming sequences, the usual considerations about promoters, polyadenylation and so forth must be applied. The promoters described above can be used, as well as tissue-specific promoters. The Tet-on and Tet-off systems are particularly useful. If the rtTA transactivator protein from the Tet-on system is placed under the control of a tissue-specific promoter, then whichever gene is placed under the control of the rtTA-regulated promoter will be activated by the addition of doxycycline, but only in certain tissues (the ones in which the transactivator is expressed). This approach has been exploited to allow the selective destruction of certain tissues to study the effects that result. In this strategy, the rtTA transactivator protein is placed under the control of an

appropriate tissue-specific promoter, and in turn the transactivator controls the expression of an intracellular toxin. In the absence of doxycycline, the toxin is not expressed. When doxycycline is added, the transactivator binds to its promoter, leading to expression of the toxin from the doxycycline-responsive promoter, but only in those cells that were expressing the transactivator. Those cells die. The technique is sometimes referred to as ablation.

Gene targeting

If it is simply sufficient that a sequence be expressed in a transgenic organism, then it will not be necessary to ensure integration of the DNA at the normal chromosomal location. For some purposes, though, it may be necessary to ensure such integration, perhaps to improve the regulation of expression of integrated DNA or to disrupt an endogenous gene. Obtaining integration at a particular site is often called gene targeting, and it can be achieved using ES cells. Targeting relies on the fact that, in ES cells, homologous recombination can take place between incoming DNA and the chromosome at the site of the corresponding sequences. The principle is that ES cells are screened after transformation for the 'correct' homologous recombination, and only those that have undergone a suitable integration event are injected into a developing embryo. Mice in which a gene has been inactivated through gene targeting are called knock-out mice. Replacing a gene with a different one is sometimes called knocking in.

Integration events may be insertional, involving a single crossover between the incoming DNA and the target gene, or replacement, involving two crossovers. Selection for acquisition of the incoming DNA can be achieved quite easily, e.g. by selection for a marker such as resistance to neomycin or G418 (Figure 9.35). However, the difficulty is that (although homologous recombination can take place) non-homologous recombination can also take place, causing sequences to be integrated at other sites in the genome. Further screening, therefore, has to be done after the initial neomycin selection for transformation, in order to identify those cells where homologous insertion has taken place. It is useful if this screening can be done selectively, as the frequency of homologous recombination may be low compared with random integration. A typical example, involving two different rounds of selection, is summarized in Figure 9.36. The incoming DNA contains two markers. Selection of cells that have acquired the first (in this case resistance to G418) identifies those that have integrated DNA. The second marker is placed so that it is lost when integration is homologous, but is likely to be retained when integration is not homologous (as incorporation of molecules by non-homologous recombination usually happens from the free ends). The second round of screening requires selection for cells that have *not* acquired the second marker, which in this case is the herpes simplex virus TK gene. Expression of this renders cells sensitive to the nucleoside analogue gancyclovir (see above), so cells

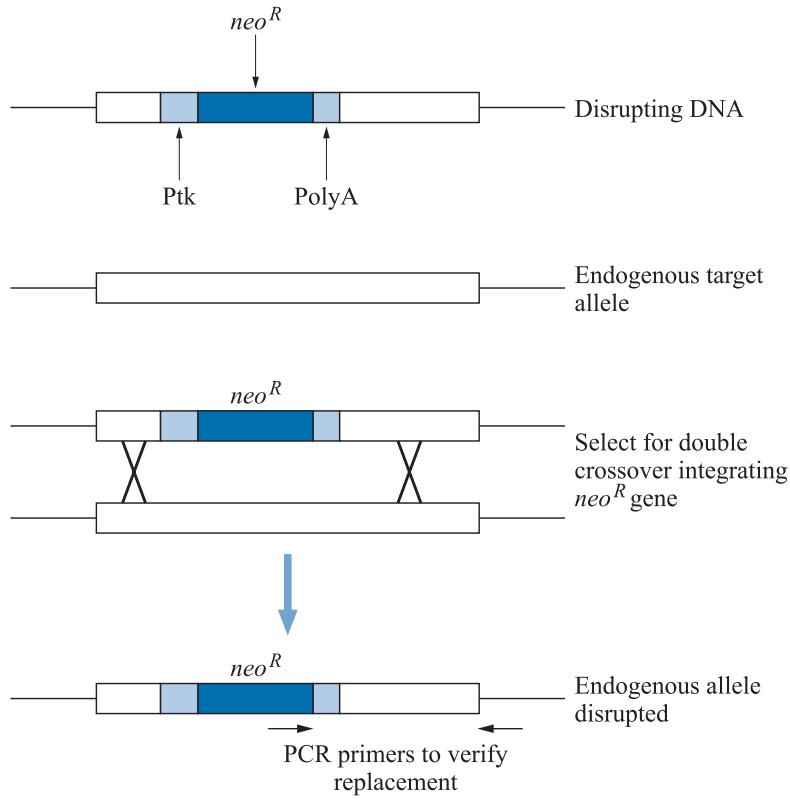


Fig 9.35 Gene disruption in an ES cell. The disrupting DNA contains a gene (*neo*^{R}) conferring resistance to neomycin or the antibiotic G418 (with a thymidine kinase promoter (Ptk) and a polyadenylation site) replacing part of the target gene. Integration of the *neo*^{R} gene at the chromosomal site of the target gene can be demonstrated by the formation of a correctly sized PCR product using appropriate primers.

that are resistant to both G418 and gancyclovir are likely to have had a targeted disruption.

Whether or not the selection against random insertion is used, it is necessary to analyse the genome of the transgenic cells to confirm that the site of integration is correct. This was done by Southern blotting in early experiments, using DNA from cultures derived from individual transformed cells. Screening many cell lines this way is laborious, and this approach has been superseded by PCR analysis. In the latter case, PCR is carried out using one primer that corresponds to a sequence within the incoming DNA and one that corresponds to a sequence close to the intended site of insertion (Figure 9.35). Homologous integration of sequences would give rise to fragments of predictable size when Southern blots were probed, or likewise, PCR products of predictable size. The cells where homologous recombination has taken place can then be taken on to generate transgenic organisms.

Sometimes it is desirable to make minor changes to a target gene, without incorporating a selectable marker at that site, in case the marker disturbs expression. One approach relies on the fact that cells that have incorporated DNA at one site are also more likely to have incorporated unlinked sequences at a different site (cotransformation). Cells are co-transformed with DNA bearing the mutated form of the target gene and with separate molecules bearing

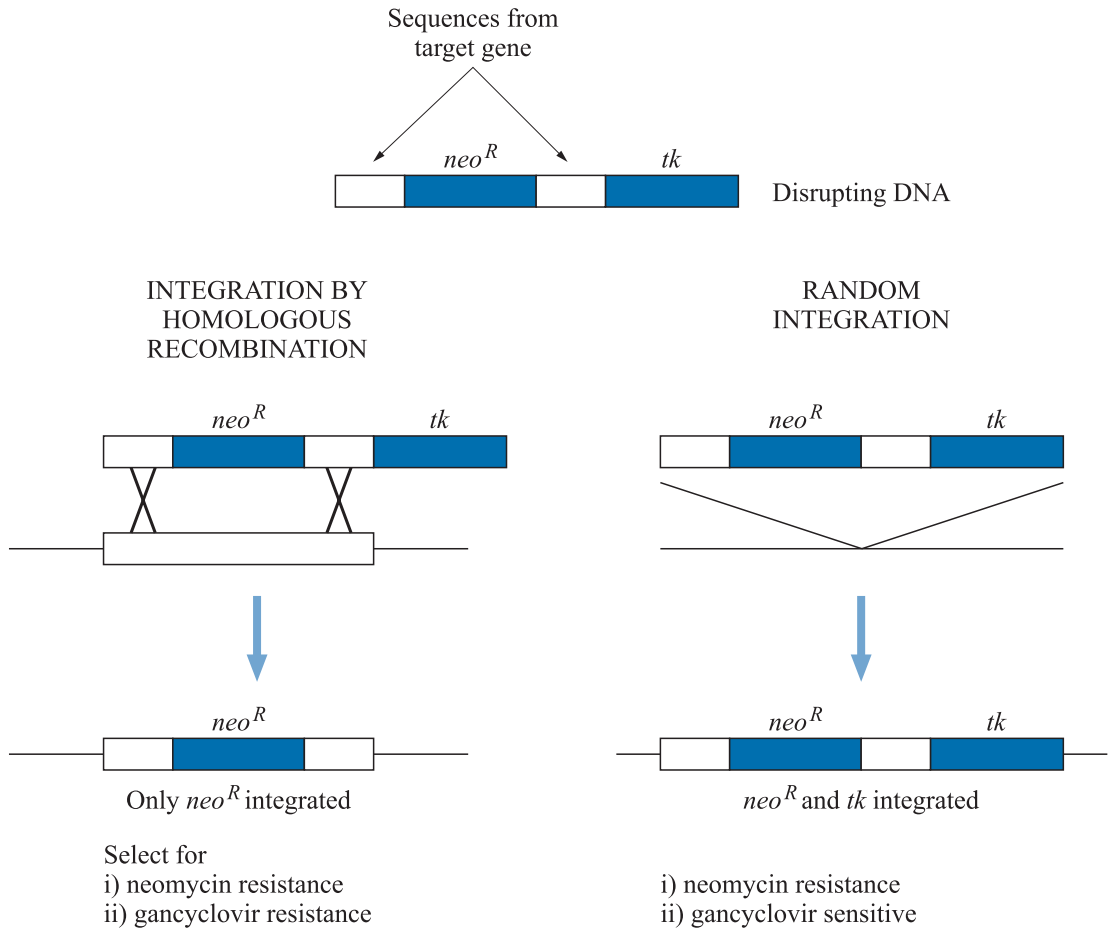


Fig 9.36 Selection for homologous recombination in gene disruption. Incoming disrupting DNA carries the *neo*^{R} gene (see Figure 9.35) and thymidine kinase (*tk*) markers. Homologous recombination between the chromosome and the sequences flanking the *neo*^{R} gene does not integrate the thymidine kinase gene; random integration (stimulated by the free end of the molecule) does.

a selectable marker. Cells that have acquired the marker by random integration are selected and then screened (e.g. by PCR) for co-acquisition of the mutation by homologous recombination at the target site.

The gene targeting approach ultimately generates individuals that are heterozygous for inactivation of the gene in question. It will often be necessary to generate homozygotes, which can be done by crossing heterozygous individuals and screening the progeny for homozygosity. (If the inactivation is a recessive lethal trait, then homozygotes will not be recovered. An elevated level of embryo mortality, resulting from the death of the homozygotes, may be observed.) Alternatively, growth of heterozygous cells in the presence of elevated concentrations of the selective agent (such as G418) can sometimes spontaneously generate homozygotes, which are presumably generated from the heterozygotes by gene conversion or a related process.

Conditional gene targeting

Conditional gene targeting can be used to bring about the controlled inactivation of a gene, and avoid some of the consequences when inactivation of a target gene is deleterious. Transgenic organisms are constructed in which the gene of interest is replaced by a copy flanked by *loxP* sites from the Cre-*lox* site-specific recombination system (see Chapter 7). The gene for the Cre recombinase is placed under a controllable promoter. Induction of this promoter leads to expression of recombinase and excision of the gene located between the *loxP* sites.

Applications

The applications of transgenic animal technology are clearly extremely broad. For example, for studies on gene regulation, transgenic animals provide a more realistic system than does the use of cultured cells. The genetic modification of economically important animals may be financially useful, either for the effects on the animals themselves or in the production of pharmaceutically important polypeptides.

Chapter 10

Examples

10.1 Introduction

The aim of this chapter is to show you how the different techniques we have covered in the preceding chapters can be put together to study biological systems. Some examples of things you might want to do will be given, and strategies will be suggested for achieving those aims. It may be helpful to read the description of the problem first, design your solution and then compare it with the suggested one. It is important to realize that there is rarely a single 'correct' answer. There may be other equally suitable strategies in addition to the ones suggested here.

Scenario 1

(a) You are studying a bacterium that grows in a particular ecological niche. You cannot culture it in the laboratory, but you can isolate small quantities of cells that microscopic analysis indicates are not contaminated with other bacteria. You want to obtain ribosomal RNA gene sequence data to study the taxonomy of the bacterium.

You could use PCR with primers to regions of ribosomal RNA genes that are conserved across a wide range of bacteria to amplify the corresponding sequences from the bacterium of interest. If the PCR product looked sufficiently specific (i.e. it appeared to be a single band in gel electrophoresis) and was in sufficient quantity, then you could determine its DNA sequence directly. If the product was not specific, then you might need to optimize the PCR, perhaps by adjusting the concentration of magnesium ions, increasing the annealing temperature or using touchdown PCR. If you obtained very little product in the PCR, then you might need to clone it before sequencing. If the polymerase you were using incorporated single A overhangs, then this might speed up the cloning process. You should be alert to the possibility of mutations occurring during the PCR amplification. The drawback with this approach is that, unless you are able to preserve some of the bacterial cells, you have no

permanent record of them. If you had enough of the bacteria to start with, then you might be able to make a genomic library in *E. coli* that could be propagated indefinitely. You could then screen the genomic library with a probe for the ribosomal RNA gene, or use PCR with DNA prepared from the library as a template. Considerations such as the genome size of the bacterium and the number of recombinants it would be convenient to maintain would determine the kind of vector used for making the genomic library.

(b) You now want to determine the complete genome sequence of the bacterium.

There are two main approaches. One is to sequence the genome systematically. This could be done by isolating a limited number of genomic clones (perhaps using BAC vectors) that would be likely to cover the entire genome, and then subcloning and sequencing each of these systematically. The second is to start by cloning small overlapping fragments at random ('shotgun cloning'), sequence and align them at the overlaps using a computer, and continue sequencing them until you have built up a complete genome. Hybrid approaches would also be possible, in which you construct a systematic collection of BAC clones, but then sequence them in a shotgun strategy. In many cloning projects, there are small regions of the genome that are not represented in the collections of overlapping sequence that you build up. It is then necessary to target these systematically, perhaps by PCR based on flanking sequence.

Scenario 2

(a) A bacterium that you can culture secretes a protein into the growth medium. Your hypothesis is that this protein is a protease. You can recover the protein from a sample of growth medium, and you already have a genomic library from the bacterium (constructed in a lambda replacement vector). You want to determine the sequence of the gene for the putative protease, to compare it with sequences of other genes for proteases. There is no complete genome sequence available for this particular bacterium.

Isolate some of the protein from a sample of growth medium, and determine partial amino acid sequence data using conventional biochemical methods. Then do one of the following:

(i) Use the sequence data to screen databases (including complete bacterial genome sequences) to identify a homologue from another organism, or organisms, to use as a probe to screen the library.

(ii) Use the amino acid sequence data to synthesize an oligonucleotide probe (probably containing mixed sites) to screen the library.

(iii) Use the oligonucleotide as a PCR primer, with the second primer derived from the library vector close to the cloning site (anchored PCR). Alternatively, if there is sufficient amino acid sequence

from elsewhere in the protein, use that to devise the second primer for PCR. Use the PCR product as a probe to screen the library.

Having identified a member of the library containing the gene, you might need to subclone it to obtain a smaller piece of genomic DNA that you could conveniently sequence.

If you did have a complete genome sequence for the bacterium of interest, then you could easily identify the gene using a computer search on the basis of the amino acid sequence determined from the protein, or you could directly use mass spectrometric analysis of the protein to identify the corresponding gene. (A full proteomic analysis of the growth medium, combined with the complete genome sequence, would allow you to obtain genes for other secreted proteins as well.)

(b) The putative protease does not show any sequence similarity to other proteases, so you want to test whether the protein really is a protease.

The simplest approach is to avoid cloning methods completely; just purify the protein and test its activity with protease substrates! If this was not feasible (perhaps if the protein is not present in large enough quantities) then you could clone the gene for the putative protease into an expression vector, express the protein and then assay it. It would be simplest to use an *E. coli* expression system, especially as we are dealing with a bacterial gene. Note that if the expressed protein did not show protease activity, that does not conclusively show it is not a protease. It could be that the protein is incorrectly folded, or needs some other protein or cofactor for activity. If the gene was not efficiently expressed in *E. coli*, then you might use the bacterial species under study for expression (assuming you had a transformation system and a suitable expression vector).

If you had a transformation system for the bacterial species under study, and methods for gene disruption were available, then you could try inactivating the gene for the putative protease. You could then test the extracellular medium from wild-type and mutant strains to see whether medium from the former had protease activity and medium from the latter did not.

(c) The bacterium turns out to be a pathogen of an agriculturally important animal species, causing liver damage. Some liver mRNAs are induced in response to infection. You want to determine their sequence, in the hope of finding out what they encode.

Isolate mRNA from livers of infected and uninfected individuals (which should otherwise be as similar as possible in their growth environment and genetic background). Make a cDNA library enriched for sequences present in the infected organisms and sequence clones at random. If microarrays were available for the animal species, then you could use those to analyse the mRNA populations from infected and uninfected organisms, without having to go through cloning and sequencing.

Scenario 3

(a) You are working on a commercially important enzyme and want to determine its three-dimensional structure by X-ray crystallography. You have a genomic clone for the enzyme.

Use an expression system to make large enough quantities of the protein to crystallize. You might use a bacterial or eukaryotic system, depending on the source of the gene and whether post-translational modification was necessary and carried out effectively in a given expression system. If the genomic clone contains introns, then you would need to isolate a cDNA clone lacking them. You should check that the expressed protein is correctly folded before using it for structure determination (e.g. by assaying for enzymatic activity).

(b) Having determined the structure, you identify a residue that may be part of the catalytic site and want to test this.

Use site-directed mutagenesis to change the codon for the residue of interest to a codon for another amino acid that should not significantly perturb the overall structure. Prepare the mutant protein using an expression system and assay its activity. Note that mutant proteins are sometimes not expressed well. For example, if the mutation perturbs the structure, then the folding and stability of the mutant protein may be impaired. If this is a problem, expression at a lower temperature may allow better folding and improve the yield of protein.

(c) Looking at the *Arabidopsis* genome sequence, you find a gene for a homologous enzyme. You want to know whereabouts in the plant the enzyme is located.

If you have a convenient assay for the enzyme, then you may not need to bother with any cloning; you can just assay different tissues for the enzyme! You can then narrow down the search by using subcellular fractionation to identify which compartment the protein is in. If, for some reason, assaying enzyme activity is not possible, then there are several other possibilities. The first looks at the protein directly but requires the generation of antibodies, which takes some time. The other approaches are more indirect, and this should be recognized in interpreting the results.

(i) Obtain a cDNA for the enzyme. With *Arabidopsis*, a suitable cDNA clone should be available from stock centres; otherwise, you should be able to use PCR to amplify the cDNA from a cDNA library. Express the cDNA to make protein, raise antibodies to the protein and use those to probe western blots of tissue extracts. Alternatively, use the antibodies to analyse tissue directly by immunocytochemistry.

(ii) Determine RNA levels in different tissues directly, by northern blots with a cDNA or genomic probe, or by qPCR. This information may already be available in published databases. Remember that the level of a given RNA does not necessarily reflect the steady-state level of the corresponding protein directly, and it will not tell you about post-translational modification of the protein, or whereabouts the protein is located within the cell.

(iii) Use reporter genes, either as a translational fusion between the reporter and the coding sequence for the enzyme or as a transcriptional fusion between the reporter and the promoter for the enzyme.

(iv) Use bioinformatic analysis of the predicted amino acid sequence of the protein to identify putative targeting sequences, and infer from those where in the cell the protein is located.

(d) Bioinformatic analysis predicts that the protein is located in the chloroplast. You would like to make radioabelled protein to see whether it can be imported into isolated chloroplasts.

You clone the cDNA into a vector allowing transcription in vitro using a suitable RNA polymerase. After transcription in vitro, the RNA is translated in vitro in the presence of one or more radioactively labelled amino acids, using a suitable protein synthesis system, such as a wheat-germ extract. The protein can then be tested for import into chloroplasts. Note that it would be important to use a full-length (or nearly full-length) cDNA for these experiments, so that any N-terminal targeting information in the protein would be likely to be intact. (Remember that you could also use a translational fusion to a reporter gene to test subcellular location, as outlined in (c) (iii) above.)

(e) You argue that an important crop plant you are studying (but for which there is no complete genome sequence) is also likely to contain the enzyme, and you want to obtain a gene for the enzyme from your crop species.

You could probe a Southern blot of genomic DNA from the plant with a genomic or cDNA clone from *Arabidopsis* to confirm whether there is a homologous gene. If you detect cross-hybridization, then you could use the *Arabidopsis* sequence as a probe for a cDNA library (and then a genomic library) from your crop species.

Alternatively, as you already have the sequence from at least two other organisms (*Arabidopsis* and the source of the enzyme we started with in (a)), you might be able to design (degenerate) PCR primers to amplify a product directly from the crop's genomic DNA (depending on the presence of introns) or RNA (i.e. cDNA), which you then sequence. Once you have obtained the right PCR product, you use it to screen a cDNA library and then a genomic library. You might also be able to get more sequence data by inverse PCR using the information you got from sequencing your initial PCR product.

(f) You want to inactivate the gene in your plant of interest, to test for its phenotype.

If you have a convenient transformation system, the most likely ways to silence the gene are using antisense RNA, or RNAi. If you do not have a convenient transformation system, then you might need to use a model plant, such as *Arabidopsis*, and study the phenotype there. For models like *Arabidopsis*, there are large collections of mutants with different genes inactivated by transposon insertion, and you may be able to obtain one of these for your gene of interest. A mutant generated in this way is likely to have the gene fully

inactivated, rather than just downregulated (which is what typically happens with antisense and RNAi).

(g) You believe the enzyme interacts with other proteins, but do not know which ones.

There are many conventional biochemical ways of approaching this. You could simply look to see whether other proteins co-purify with the protein of interest using a range of fractionation techniques. If you have already made antibodies to your enzyme of interest, you could try immunoprecipitation from extracts using the antibodies and see whether other proteins are consistently immunoprecipitated at the same time. The likelihood of recovering an interacting protein in co-purification or immunoprecipitation could be increased by treatment with cross-linking agents, to stabilize the interactions. (This requires careful titration of the amount of crosslinker, though.) Alternatively, you could try two-hybrid screening of a cDNA library using the coding sequence of your enzyme as bait. You isolate cDNAs whose products interact with the bait in the two-hybrid screen, sequence them and try to identify what they encode using database searches.

Scenario 4

(a) You are studying a human genetic disease that causes benign growths in the brain, kidneys, skin, heart and lungs. You have several families whose pedigrees indicate a single-gene mutation, and the site of the mutation appears to map between two closely linked genetic markers that have already been located on the genome.

Analysis *in silico* of human genome sequence data should allow you to identify likely coding regions in the area of interest. Microarray data should indicate which are expressed in the tissues affected. You could then look for a consistent association of mutation in one of the candidate sequences with the disorder. PCR of target regions would be a convenient way of doing so.

To verify that you have found the right gene, you could knock out the corresponding gene in mice to look for a similar phenotype. This could be done by targeted disruption in embryonic stem cells, followed by generation of heterozygotes and crossing to generate homozygotes.

(b) You find that no homozygous knockout mice can be recovered, as embryos die early in development. Nevertheless, you want to determine the effect of homozygous gene knockout on astrocytes, a particular cell type, in mice.

You generate two homozygous lines of transgenic mice. One is modified to contain a suitable region of the gene of interest flanked by *loxP* sites (probably by inserting them into introns), and should have normal expression of the gene. The second transgenic line contains a gene for Cre recombinase under the control of an astrocyte-specific promoter. Crossing the lines generates intermediate heterozygotes, and intercrossing these generates (among others) lines that are homozygous for the *loxP*-flanked gene and homozygous (or heterozygous) for

the recombinase. The recombinase gene will be expressed in astrocytes, as it is under the control of an astrocyte-specific promoter. This leads to deletion within both alleles of the target gene, and thus individuals that are homozygous for a knockout in astrocytes. This can be verified by PCR. Note that the germ cells of the intermediate heterozygotes that were crossed to make homozygotes will retain the intact, but *loxP*-flanked, target gene. This is because the recombinase is not expressed in the germ cells, and these cells, therefore, do not have the *loxP*-mediated deletion. (This is a complex experimental strategy. More details can be found in Chapter 7 and in the paper on which this example was based (Uhlmann *et al.*, 2002.)

References

This is not meant to be an exhaustive list of references in support of the material in the book. It is simply a collection of articles and books to offer further information on the topics discussed. Many of the companies marketing enzymes and other reagents for recombinant DNA work (such as Boehringer, Invitrogen, Novagen, Promega and Stratagene) provide very detailed and informative descriptions of vectors and techniques in their catalogues. There are also a number of very helpful laboratory manuals, including the two below. They are relevant to many of the chapters, so are listed separately first, rather than for each chapter.

Ausubel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A. & Struhl, K. eds. (2006). *Current Protocols in Molecular Biology*. John Wiley.

Sambrook, J. & Russell, D. W. (2001). *Molecular Cloning: A Laboratory Manual*. 3rd edition. Cold Spring Harbor Laboratory Press.

Chapter 1

Kessler, C. & Manta, V. (1990). Specificity of restriction endonucleases and DNA modification methyltransferases. *Gene*, **92**, 1–248.

Stekel, D. (2003). *Microarray Bioinformatics*. Cambridge University Press.

Wilson, G. G. & Murray, N. E. (1991). Restriction and modification systems. *Annual Reviews of Genetics*, **25**, 585–627.

Chapter 2

Bower, M. A., Spencer, M., Matsumura, S., Nisbet, R. E. R. & Howe, C. J. (2005). How many clones need to be sequenced from a single forensic or ancient DNA sample in order to determine a reliable consensus sequence? *Nucleic Acids Research*, **33**, 2549–2556.

Cooper, A. & Poinar, H. N. (2000). Ancient DNA: do it right or not at all. *Science*, **289**, 1139.

Hagelberg, E., Gray, I. C. & Jeffreys, A. J. (1991). Identification of the skeletal remains of a murder victim by DNA analysis. *Nature*, **352**, 427–429.

McPherson, M. J. & Møller, S. G. (2006). *PCR: The Basics*. Taylor & Francis.

Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B. & Erlich, H. A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, **239**, 487–491.

Suzuki, R., Yoshikawa, T., Ihira, M., Enomoto, Y., Inagaki, S., Matsumoto, K., Kato, K., Kudo, K., Kojima, S. & Asano, Y. (2006). Development of the loop-mediated isothermal amplification method for rapid detection of cytomegalovirus DNA. *Journal of Virological Methods*, **132**, 216–221.

Thomas, R. H., Schaffner, W., Wilson, A. C. & Pääbo, S. (1989). DNA phylogeny of the extinct marsupial wolf. *Nature*, **340**, 465–467.

Chapter 3

Hanahan, D. (1983). Studies on transformation of *Escherichia coli* with plasmids. *Journal of Molecular Biology*, **166**, 557–580.

Hanahan, D., Jessee, J. & Bloom, F. R. (1991). Plasmid transformation of *Escherichia coli* and other bacteria. *Methods in Enzymology*, **204**, 63–113.

Knoblauch, M., Hibberd, J.M., Gray, J.C. & van Bel., A.J.E. (1999). A galinstan expansion femtosyringe for microinjection of eukaryotic organelles and prokaryotes. *Nature Biotechnology*, **17**, 906–909.

Summers, D.K. (1996). *The Biology of Plasmids*. Blackwell Science.

Wilson, G.G. & Murray, N.E. (1991). Restriction and modification systems. *Annual Reviews of Genetics*, **25**, 585–627.

Chapter 4

Clackson, T. & Lowman, H.B. eds (2004). *Phage Display: A Practical Approach*. Oxford University Press.

Groisman, E.A., Castilho, B.A. & Casadaban, M.J. (1984). In vivo DNA cloning and adjacent gene fusing with a mini-Mu-*lac* bacteriophage containing a plasmid replicon. *Proceedings of the National Academy of Sciences, USA*, **81**, 1480–1483.

Ioannou, P.A., Amemiya, C.T., Garnes, J., Kroisel, P.M., Shizuya, H., Chen, C., Batzer, M.A. & de Jong, P.J. (1994). A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. *Nature Genetics*, **6**, 84–89.

Kaiser, K., Murray, N.E. & Whittaker P.A. (1995). Construction of representative genomic DNA libraries using phage lambda replacement vectors. In D.M. Glover and B.D. Hames eds, *DNA Cloning 1: A Practical Approach*. IRL Press, pp. 37–84.

Kim, U.-J., Birren, B.W., Slepak, T., Mancino, V., Boysen, C., Kang, H.-L., Simon, M.I. & Shizuya, H. (1996). Construction and characterization of a human bacterial artificial chromosome library. *Genomics*, **34**, 213–218.

Ptashne, M. (2004). *A Genetic Switch: Phage Lambda Revisited*. Cold Spring Harbor Laboratory Press.

Short, J.M., Fernandez, J.M., Sorge, J.A. & Huse, W.D. (1988). Lambda ZAP: a bacteriophage lambda expression vector with *in vivo* excision properties. *Nucleic Acids Research*, **16**, 7583–7600.

Sternberg, N.L. (1992). Cloning high molecular weight DNA fragments by the bascteriophage P1 system. *Trends in Genetics*, **8**, 11–16.

Chapter 5

Frohman, M.A., Dush, M.K. & Martin, G.R. (1988). Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proceedings of the National Academy of Sciences, USA*, **85**, 8998–9002.

Soares, M.B., Bonaldo, M. de F., Jelene, P., Su, L., Lawton, L. & Efstratiadis, A. (1994). Construction and characterization of a normalized cDNA library. *Proceedings of the National Academy of Sciences*, *USA*, **91**, 9228–9232.

Chapter 6

Clackson, T. & Lowman, H.B. eds (2004). *Phage Display: A Practical Approach*. Oxford University Press.

Fields, S. & Sternglanz, R. (1994). The two-hybrid system: an assay for protein–protein interactions. *Trends in Genetics*. **10**, 286–292.

Hediger, M.A., Coady, M.J., Ikeda, T.S. & Wright, E.M. (1987). Expression, cloning and cDNA sequencing of the Na^+/glucose co-transporter. *Nature*, **330**, 379–381.

Jaeger, S., Eriani, G. & Martin, F. (2004). Results and prospects of the yeast three-hybrid system. *FEBS Letters*, **556**, 7–12.

Schaffitzel, C., Hanes, J., Jermutus, L. & Plückthun, A. (1999). Ribosome display: an in vitro method for selection and evolution of antibodies from libraries. *Journal of Immunological Methods*. **231**, 119–135.

SenGupta, D.J., Zhang, B., Kraemer, B., Pochart, P., Fields, S. & Wickens, M. (1996). A three-hybrid system to detect RNA–protein interactions in vitro. *Proceedings of the National Academy of Sciences, USA*, **93**, 8496–8501.

Chapter 7

Baulcombe, D. (2005). RNA silencing. *Trends in Biochemical Sciences*, **30**, 290–293.

Hamilton, A.J. & Baulcombe, D.C. (1999). A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science*, **286**, 950–952.

Hammond, S.M. (2005). Dicing and slicing: the core machinery of the RNA interference pathway. *FEBS Letters*, **579**, 5822–5829.

Kolmar, H. & Fritz, H.-J. (1995). Oligonucleotide-directed mutagenesis with single-stranded cloning vectors. In D.M. Glover and B.D. Hames eds, *DNA Cloning 1: A Practical Approach*. IRL Press, pp. 193–224.

Rothstein, R. (1991). Targeting, disruption, replacement and allele rescue: integrative DNA transformation in yeast. *Methods in Enzymology*, **194**, 281–301.

Timmons, L. & Fire, A. (1998). Specific interference by ingested dsRNA. *Nature*, **395**, 854.

Tyagi, R., Lai, R. & Duggleby, R.G. (2004). A new approach to 'megaprimer' polymerase chain reaction mutagenesis without an intermediate gel purification step. *BMC Biotechnology*, **4**, 2.

Wang, W.Y. & Malcolm, B.A. (1999). Two-stage PCR protocol allowing introduction of multiple mutations, deletions and insertions using QuikChange™ site-directed mutagenesis. *Biotechniques*, **26**, 680–682.

Yem, L., Svendsen, J., Lee, J.-S., Gray, J.T., Magnier, M., Baba, T., D'Amato, R.J. & Mulligan, R.C. (2004). Exogenous control of mammalian gene expression through modulation of RNA self-cleavage. *Nature*, **431**, 471–476.

Chapter 8

Brosius, J., Erfle, M. & Storella, J. (1985). Spacing of the −10 and −35 regions in the *tac* promoter. *Journal of Biological Chemistry*, **260**, 3539–3541.

Brosius, J. & Holy, A. (1984). Regulation of ribosomal RNA promoters with a synthetic *lac* operator. *Proceedings of the National Academy of Sciences, USA*, **81**, 6929–6933.

Chong, S., Mersha, F.B., Comb, D.G., Scott, M.E., Landry, D., Vence, L.M., Perler, F.B., Benner, J., Kucera, R.B., Hirvonen, C.A., Pelletier, J.J., Paulus, H. & Xu, M.-Q. (1997). Single-column purification of free recombinant proteins using a self-cleavable affinity tag derived from a protein-splicing element. *Gene*, **192**, 271–281.

Chubet, R.G. & Brizzard, B.L. (1996). Vectors for expression and secretion of FLAG epitope-tagged proteins in mammalian cells. *Biotechniques*, **20**, 136–141.

Finn, R.D., Kapelioukh, I. & Paine, M.J.I. (2005). Rainbow tags: a visual tag system for recombinant protein expression and purification. *Biotechniques*, **38**, 387–392.

Friedel, R.H., Plump, A., Lu, X., Spilker, K., Joliocoeur, C., Wong, K., Venkatesh, T.R., Yaron, A., Hynes, M., Chen, B., Okada, A., McConnell, S.K., Rayburn, H. & Tessier-Lavigne, M. (2005). Gene targeting using a promoterless gene trap vector (''targeted trapping'') is an efficient method to mutate a large fraction of genes. *Proceedings of the National Academy of Sciences, USA*, **102**, 13 188–13 193.

Glover, D.M. & Hames, B.D. eds (1995). *DNA Cloning 2, Expression Systems: A Practical Approach*. IRL Press.

Johnson, A.A.T., Hibberd, J.M., Gay, C., Essah, P.A., Haseloff, J., Tester, M. & Guiderdoni, E. (2005). Spatial control of transgene expression in rice (*Oryza sativa* L.) using the GAL4 enhancer trapping system. *Plant Journal*, **41**, 779–789.

Lilius, G., Persson, M., Buflow, L. & Mosbach, K. (1991). Metal affinity precipitation of proteins carrying genetically attached polyhistidine affinity tails. *European Journal of Biochemistry*, **198**, 499–504.

Chapter 9

Bacteria

Cabello, F.C., Sartakova, M.L. & Dobrikova, E.Y. (2001). Genetic manipulation of spirochetes – light at the end of the tunnel. *Trends in Microbiology*, **9**, 245–248.

Hanahan, D., Jessee, J. & Bloom, F.R. (1991). Plasmid transformation of *Escherichia coli* and other bacteria. *Methods in Enzymology*, **204**, 63–113.

Kieser, T., Bibb, M.J., Buttner, M.J., Chater, K.F. & Hopwood, D.A. (2000). *Practical* Streptomyces *genetics*. John Innes Foundation.

Kunji, E.R.S., Slotboom, D.-J. & Poolman, B. (2003). *Lactococcus lactis* as host for overproduction of functional membrane proteins. *Biochimica et Biophysica Acta*, **1610**, 97–108.

Le Grice, S.F.J. (1990). Regulated promoter for high-level expression of heterologous genes in *Bacillus subtilis*. *Methods in Enzymology*, **185**, 201–214.

Marx, C.J. & Lidstrom, M.E. (2001). Development of improved versatile broad-host-range vectors for use in methylotrophs and other Gram-negative bacteria. *Microbiology*, **147**, 2065–2075.

Nagarajan, V. (1990). System for secretion of heterologous proteins in *Bacillus subtilis*. *Methods in Enzymology*, **185**, 214–228.

Scott, H.N., Laible, P.D. & Hanson, D.K. (2003). Sequences of versatile broad-host-range vectors of the RK2 family. *Plasmid*, **50**, 74–79.

Fungi

Cereghino, J.L. & Cregg, J.M. (2000). Heterologous protein expression in the methylotrophic yeast *Pichia pastoris*. *FEMS Microbiology Reviews*, **24**, 45–66.

Funk, M., Niedenthal, R., Mumberg, D., Brinkmann, K., Rönicke, V. & Henkel, T. (2002). Vector systems for heterologous expression of proteins in *Saccharomyces cerevisiae*. *Methods in Enzymology*, **350**, 248–257.

Gouka, R.J., Punt, P.J. & van den Hondel, C.A.M.J.J. (1997). Efficient production of secreted proteins by *Aspergillus*: progress, limitations and prospects. *Applied Microbiology and Biotechnology*, **47**, 1–11.

Ruiz-Diez, B. (2002). Strategies for the transformation of filamentous fungi. *Journal of Applied Microbiology*, **92**, 189–195.

Schlessinger, D. (1990). Yeast artifical chromosomes: tools for mapping and analysis of complex genomes. *Trends in Genetics*, **6**, 248–258.

Siam, R., Dolan, W.P. & Forsburg, S.L. (2004). Choosing and using *Schizosaccharomyces pombe* plasmids. *Methods*, **33**, 189–198.

Song, H., Li, Y., Fang, W., Geng, Y., Wang, X., Wang, M. & Qiu, B. (2003). Development of a set of expression vectors in *Hansenula polymorpha*. *Biotechnology Letters*, **25**, 1999–2006.

Toews, M.W., Warmbold, J., Konzack, S., Rischitor, P., Veith, D., Vienken, K., Vinuesa, C., Wei, H. & Fischer, R. (2004). Establishment of mRFP1 as a fluorescent marker in *Aspergillus nidulans* and construction of expression vectors for high-throughput protein tagging using recombination in vitro (GATEWAY). *Current Genetics*, **45**, 383–389.

Wang, L., Kao, R., Ivey, F.D. & Hoffmann, C.S. (2004). Strategies for gene disruptions and plasmid constructions in fission yeast. *Methods*, **33**, 199–205.

Wang, Y.-C.M., Chuang, L.L., Lee, F.W.F. & Da Silva, N.A. (2003). Sequential cloned gene integration in the yeast *Kluyveromyces lactis*. *Applied Microbiology and Biotechnology*, **62**, 523–527.

Algae and plants

Bateman, J.M. & Purton, S. (2000). Tools for chloroplast transformation in *Chlamydomonas*: expression vectors and a new dominant selectable marker. *Molecular and General Genetics*, **263**, 404–410.

Castle, L.A., Siehl, D.L., Gorton, R., Patten, P.A., Chen, Y.H., Bertain, S., Cho, H.-J., Duck, N., Wong, J., Liu, D. & Lassner, M.W. (2004). Discovery and directed evolution of a glyphosate tolerance gene. *Science*, **304**, 1151–1154.

Clough, S.J. & Bent, A.F. (1998). Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *The Plant Journal* **16**, 735–743.

Hare, P.D. & Chua, N.-H. (2002). Excision of selectable markers from transgenic plants. *Nature Biotechnology*, **20**, 575–580.

Hellens, R.P., Edwards, E.A., Leyland, N.R., Bean, S. & Mullineaux, P.M. (2000). pGreen: a versatile and flexible binary Ti vector for *Agrobacterium*-mediated plant transformation. *Plant Molecular Biology*, **42**, 819–832.

Ma, J.K.C., Drake, P.M.W. & Christou, P. (2003). The production of recombinant pharmaceutical proteins in plants. *Nature Reviews Genetics*, **4**, 794–805.

Ogita, S., Uefuji, H., Yamaguchi, Y., Koizumi, N. & Sano, H. (2003). Producing decaffeinated coffee plants. *Nature*, **423**, 823.

Walker, T.L., Collet, C. & Purton, S. (2005). Algal transgenics in the genomic era. *Journal of Phycology*, **41**, 1077–1093.

Weigel, D. & Glazebrook, J. (2002). Arabidopsis: *A Laboratory Manual*. Cold Spring Harbor Laboratory press.

Wright, M., Dawson, J., Dunder, E., Suttie, J., Reed, J., Kramer, C., Chang, Y., Novitzky, R., Wang, H. & Artim-Moore, L. (2001). Efficient biolistic transformation of maize (*Zea mays* L.) and wheat (*Triticum aestivum* L.) using the phosphomannose isomerase gene, *pmi*, as the selectable marker. *Plant Cell Reports*, **20**, 429–436.

Organelles

Bonnefoy, N. & Fox, T.D. (2001). Genetic transformation of *Saccharomyces cerevisiae* mitochondria. *Methods in Cell Biology*, **65**, 381–396.

Lutz, K.A., Bosacchi, M.H. & Maliga, P. (2006). Plastid marker-gene excision by transiently expressed CRE recombinase. *Plant Journal*, **45**, 447–456.

Maliga, P. (2004). Plastid transformation in higher plants. *Annual Review of Plant Biology*, **55**, 289–313.

Mireau, H., Arnal, N. & Fox, T. D. (2003). Expression of Barstar as a selectable marker in yeast mitochondria. *Molecular Genetics and Genomics*, **270**, 1–8.

Caenorhabditis elegans

Berezikov, E., Bargmann, C. I. & Plasterk, R. H. A. (2004). Homologous gene targeting in *Caenorhabditis elegans* by biolistic transformation. *Nucleic Acids Research*, **32**, e40.

Praitis, V., Casey, E., Collar, D. & Austin, J. (2001). Creation of low-copy integrated transgenic lines in *Caenorhabditis elegans*. *Genetics*, **157**, 1217–1226.

Timmons, L. & Fire, A. (1998). Specific interference by ingested dsRNA. *Nature*, **395**, 854.

Insects

Liepman, A. H., Wilkerson, C. G. & Keegstra, K. (2005). Expression of cellulose synthase-like (Csl) genes in insect cells reveals that CslA family members encode mannan synthases. *Proceedings of the National Academy of Sciences, USA*, **102**, 2221–2226.

Luckow, V. A., Lee, S. C., Barry, G. F. & Olins, P. O. (1993). Efficient generation of infectious recombinant baculoviruses by site-specific transposon-mediated insertion of foreign genes into a baculovirus genome propagated in *Escherichia coli*. *Journal of Virology*, **67**, 4566–4579.

Richardson, C. D. ed. (2005). *Baculovirus Expression Protocols*. Humana Press.

Rong, Y. S. & Golic, K. G. (2000). Gene targeting by homologous recombination in *Drosophila*. *Science*, **288**, 2013–2018.

Schotta, G. & Reuter, G. (2000). Controlled expression of tagged proteins in *Drosophila* using a new modular P-element vector system. *Molecular and General Genetics*, **262**, 916–920.

Venken, K. J. T. & Bellen, H. J. (2005). Emerging technologies for gene manipulation in *Drosophila melanogaster*. *Nature Reviews Genetics*, **6**, 167–178.

Wimmer, E. A. (2003). Applications of insect transgenesis. *Nature Reviews Genetics*, **4**, 225–232.

Mammals

Basu, J., Compitello, G., Stromberg, G., Willard, H. F. & van Bokkelen, G. (2005). Efficient assembly of *de novo* human artificial chromosomes from large genomic loci. *BMC Biotechnology*, **5**, 21.

Basu, J. & Willard, H. F. (2005). Artificial and engineered chromosomes: non-integrating vectors for gene therapy. *Trends in Molecular Medicine*, **11**, 251–258.

Brocard, J., Warot, X., Wendling, O., Messaddeq, N., Vonesch, J.-L. Chambon, P. & Metzger, D. (1997). Spatio-temporally controlled site-specific somatic mutagenesis in the mouse. *Proceedings of the National Academy of Sciences, USA*, **94**, 14 559–14 563.

Gordon, J. W. (1993). Production of transgenic mice. *Methods in Enzymology*, **225**, 747–753.

Grimm, S. (2004). The art and design of genetic screens: mammalian culture cells. *Nature Reviews Genetics*, **5**, 179–189.

Kistner, A., Gossen, M., Zimmerman, F., Jerecic, J., Ullmer, C., Lübbert, H. & Bujard, H. (1996). Doxycycline-mediated quantitative and tissue-specific control of gene expression in transgenic mice. *Proceedings of the National Academy of Sciences*, USA, **93**, 10 933–10 938.

Verma, I.M. & Weitzmann, M.D. (2005). Gene therapy: twenty-first century medicine. *Annual Review of Biochemistry*, **74**, 711–738.

Chapter 10

Uhlmann, E.J., Wong, M., Baldwin, R.L., Bajenaru, M.L., Onda, H., Kwiatkowski, D.J., Yamada, K. & Gutmann, D.H. (2002). Astrocyte-specific *TSC1* conditional knockout mice exhibit abnormal neuronal organization and seizures. *Annals of Neurology*, **52**, 285–296.

Index